Genome Biology

METHOD

Open Access

DAGIP: alleviating cell-free DNA sequencing biases with optimal transport

Antoine Passemiers[1*], Stefania Tuveri[2], Tatjana Jatsenko[2], Adriaan Vanderstichele[3,4], Pieter Busschaert[3,4], An Coosemans[7], Dirk Timmerman[3], Sabine Tejpar[8], Peter Vandenberghe[9,10], Diether Lambrechts[5,6], Daniele Raimondi[1,11], Joris Robert Vermeesch[2] and Yves Moreau[1]

*Correspondence: antoine.passemiers@gmail.com

[1] Dynamical Systems, Signal Processing and Data Analytics (STADIUS), KU Leuven, Leuven, Belgium
Full list of author information is available at the end of the article

Abstract

Cell-free DNA (cfDNA) is a rich source of biomarkers for various pathophysiological conditions. Preanalytical variables, such as the library preparation protocol or sequencing platform, are major confounders of cfDNA analysis. We present DAGIP, a novel data correction method that builds on optimal transport theory and deep learning, which explicitly corrects for the effect of such preanalytical variables and can infer technical biases. Our method improves cancer detection and copy number alteration analysis by alleviating the sources of variation that are not of biological origin. It also enhances fragmentomic analysis of cfDNA. DAGIP allows the integration of cohorts from different studies.

Background

Cell-free DNA (cfDNA) has been identified as a promising source of biomarkers for the detection of fetal aneuploidy [1, 2], transplant rejection [3], incipient tumors [4], autoimmune disease [5], or inflammatory disease [6]. While cfDNA fragments in healthy individuals primarily originate from the apoptotic release of DNA from cells of hematopoietic origin [7], these fragments can also be of tumoral origin in cancer patients. While most clinical applications of cfDNA in oncology focus on finding tumor mutations (e.g., using a targeted panel of cancer driver variants) [8, 9], a lot of research has been carried out around the analysis of coverage and fragmentome profiles. The somatic copy number aberrations (CNAs) carried by the genome of cancerous cells are detectable by low-coverage whole-genome sequencing and downstream analysis of coverage profiles from cancer patients [10, 11]. Fragmentomic analysis of cfDNA offers the possibility to detect new sensitive biomarkers for cancer detection [12, 13], as cfDNA fragments mirror the chromatin accessibility, nucleosome positioning, and degradation pattern of their tissue of origin [14–17]. For this reason, CNA calling can be complemented with fragmentation profile analysis based on fragment length, as well as

positional information [12, 18–20]. Circulating tumor-derived DNA (ctDNA) fragments have been found to be typically shorter [14, 21]. Moreover, cfDNA fragment end motifs are non-random, and their frequencies have been shown to be altered in cancer patients due to changes in fragmentation patterns [22]. Jagged end size was observed to be higher in cancer and fetal cfDNA [23]. Finally, all these different properties of cfDNA have been demonstrated to be related to the activities of the *DFFB*, DNASE1L3, and DNASE1 endonucleases [24–26]. Beyond fragmentomics, methylation patterns are indicative of the tissue of origin, and methylation signatures have been exploited for sensitive cancer detection and tissue-of-origin identification [27]. Finally, recent work has been devoted toward integrating multiple properties of cfDNA within a single multimodal analysis approach, including variant calling, CNAs, methylation, and fragmentomic profiles, as well as other complementary sources of information such as nucleosome-depleted region (NDR) profiles [28] or fusion gene detection [29]. Because cfDNA can be collected in a non- or minimally-invasive manner (e.g., blood draw), and as a result of the cost-effectiveness of shallow whole-genome sequencing, liquid biopsies are a valuable candidate for population-wide cancer screening [4, 7, 30] and diagnosis, and considerable research has been devoted to assessing their clinical utility [31].

However, the development of reliable models that are predictive of relevant clinical outcomes (e.g., diagnosis) remains challenging because of the limited number of available cases (especially for disorders with smaller incidence rate), the high dimensionality of cfDNA data, and the various sources of biases related to preanalytical settings. These latter biases mainly arise when protocol changes are introduced over time or between different centers. For example, the choice of blood collection tube might affect cfDNA concentrations and the prominence of leukocyte DNA in plasma samples [32, 33], which could in turn affect the detection of low-frequency variations originating from cancerous cells. Other preanalytical factors include the delay before centrifugation, protocols for plasma separation, and plasma storage conditions [34]. For example, two-step centrifugation reduces contamination by genomic DNA due to reduced white blood cell lysis, compared to one-step centrifugation [35]. Moreover, some DNA extraction platforms, such as Maxwell and QIAsymphony, preferentially isolate short fragments over long ones [36]. The choice of library preparation kit directly affects the distribution of read counts, as the polymerase enzymes used in these kits have different levels of efficiency in amplifying fragments with low vs. high GC-content [37]. For instance, some library preparation kits (e.g., Nextera XT) introduce a bias toward low-GC regions [38]. Multiplexed sequencing without suitable dual indexing can result in barcode swapping, and the swapping rates are sequencing platform-dependent (e.g., higher on HiSeqX or 4000 compared to MiSeq) [39]. Index swapping mechanism is caused both by multiplex PCR and flow cell chemistry and is responsible for cross-contamination within the same pool [40]. Finally, the choice of sequencing instrument itself also plays a role. For example, different GC-content bias profiles have been reported for Illumina MiSeq and NextSeq platforms, compared to PacBio or HiSeq [41].

The aforementioned preanalytical settings can affect the sequencing outcomes and potentially invalidate direct statistical analyses on the resulting data. Moreover, these distributional shifts [42] are not properly handled by classical machine learning algorithms and are responsible for performance drops on test sets. Mitigating these biases

is therefore of utmost importance in strengthening biological signals and guaranteeing detection performance in independent cohorts. Such a task typically falls in the category of domain adaptation (DA) [43] problems in the machine learning literature. In DA settings, a statistical model is tested on data produced under different measurement conditions (i.e., domains) than the ones on which the model has been trained, and specific algorithms are designed to account for these differences. When multiple source domains coexist, the problem is referred to as a domain generalization problem [44]. Previous work on DA includes the following. Domain adversarial learning [45] aims at constructing a common representation space for all domains and relies on gradient reversal for enforcing domains to be indistinguishable in that latent space [45, 46]. Domain separation networks [46] is a reconstruction-based method consisting in learning both common and domain-specific representations. On the other hand, Cycle-Consistent Adversarial DA [47] reconstructs samples from the target domain using samples from the source domain as input. Adversarial methods have also been complemented with metric learning [48] and data synthesis in the target domain [49]. Discriminator-free domain adversarial learning [50] works similarly but does not rely on a neural network classifier for predicting the variable of interest. Distance-based methods rely on Maximum Mean Discrepancy [51], Central Moment Discrepancy (CMD) [52], or higher-order moment matching [53] to encourage the moments of the empirical data distributions to coincide. Finally, other categories of DA methods include information-theoretic DA [54] and optimal transport (OT) [55–58]. It should be noted that most existing DA methods use a latent space to represent the samples, which means that the debiased representation is not directly interpretable, which runs against the ubiquitous need for interpretability and explainability in human genetics [59]. A key motivation for our work is thus to design a DA method that adjusts cfDNA profiles in a transparently interpretable manner, by operating in the original space (i.e., without having recourse to a latent space as domain adversarial methods would) and preserving the quantiles in the original data (e.g., log-ratios, *z*-scores).

Previous work on the bias correction of copy-number profiles has mostly been directed toward GC-content and mappability bias correction [29, 60–64]. Distance learning and *k*-nearest neighbors have also been proposed [65] to correct coverage profiles. As opposed to previous work, the latter approach exploits information from the whole data set to correct each individual sample. `dryclean` [66] uses online robust principal component analysis (rPCA) to isolate foreground biological signals from background technical artifacts. Finally, similarly to `dryclean`, tangent and pseudo-tangent normalization [67] methods were proposed for reducing technical noise in coverage profiles and improving the detection of somatic CNAs.

In this work, we focus on the bias correction of genome-wide copy-number profiles (i.e., coverage profiles [4]), as well as cfDNA fragmentomics modalities such as end motif frequencies or nucleosome positioning. Our method, coined DAGIP, builds on OT theory [68, 69], which is itself based on strong mathematical bases and allows to define a sample-to-sample relationship across wet-lab protocols in a highly interpretable manner, as samples can be corrected in the original data space (e.g., fragment size frequencies) directly. In summary, we aim at correcting samples from a source domain (a given wet-lab protocol) towards a target domain (a different protocol) to

enable more robust downstream analysis in the latter domain. As the ultimate goal is to go beyond the classical case–control setting and build models capable of accurately processing data from various sources, we hypothesized that bias correction is a good candidate to increase the effective size of available data sets through their fusion, and allow benefiting from the scalability of machine learning models for enhanced performance. This flexibility would, among other things, reduce the need for laboratories to consistently build new reference sets, as well as enable high reusability of older samples or data collected in unrelated studies. In this work, we aim at reducing non-biological variance and enhance cancer detection, while preserving the original biological signals (e.g., copy number aberrations) in each analyzed sample.

Results

In Table 1, we summarized the different data sets and preanalytical settings used in our study. Throughout the paper, we referred to *domain* as a wet-lab protocol or set of preanalytical settings. The terms *domain* and *protocol* were used interchangeably. Each *data set* included multiple domains and was dedicated to specific experiments or to the detection of a specific cancer type.

Our methods are illustrated in Fig. 1. Given two groups of sample-derived profiles (e.g., coverage profiles, fragment size profiles) from wet-lab protocols 2 and 1, structured as matrices X and Y, respectively, our bias correction approach (coined DAGIP) relies on a neural network model to explicitly estimate the bias of sample i from its profile $X_{i\cdot}$ (Fig. 1A). Ideally, the two groups should be matched and therefore be representative of the same population, and this population should be as large as possible to guarantee the reliability of the model. The method first computes pairwise distances between sample-derived profiles and solves the associated OT problem. The solution to this problem is referred to as the transport plan, defines sample-to-sample similarities, and is used to guide the direction of the correction function (Fig. 1B). The outputs of the algorithm are a corrected version of X denoted by $\mathcal{X}$, and a trained model which can be used to correct any new sample processed with protocol 2 independently, without the need to process matched samples with protocol 1 (Fig. 1C). Algorithmic details are provided in Methods section.

We used different methods to validate the relevance of bias correction. A direct and principled approach consists in processing technical replicates using different protocols and evaluating whether bias correction enables the identification of sample pairs. For this purpose, 64 ovarian carcinoma cases have been processed with different wet-lab settings (OV data set), as well as 563 pregnancy samples collected during routine Non-Invasive Prenatal Testing (NIPT data set). Next, we evaluated whether bias correction retains biological information from coverage profiles, by performing CNA analysis using `ichorCNA` in different settings and showing the concordance of called CNAs. Finally, we evaluated the overall relevance of bias correction through the enhancement in cancer detection. For this purpose, we devised a specific cross-validation scheme for machine learning, depicted later in Fig. 8A–B.

Table 1 Summary of the data sets used in this study

Data set	Condition/ setting	Domain	Size	Library preparation kit	Index type	Paired-end?	Sequencer
NIPT	Pregnancy (kit validation)	$\mathcal{D}_{1,a}$	$2 \times 66^* \begin{cases} 66 \\ 66 \end{cases}$	TruSeq Nano	TruSeq Nano	No	HiSeq 4000
		$\mathcal{D}_{1,b}$		Kapa Hyper-Prep	Kapa dual	No	HiSeq 4000
	Pregnancy (adapter validation)	$\mathcal{D}_{2,a}$	$2 \times 179^* \begin{cases} 179 \\ 179 \end{cases}$	Kapa Hyper-Prep	IDT	No	HiSeq 4000
		$\mathcal{D}_{2,b}$		Kapa Hyper-Prep	Kapa dual	No	HiSeq 4000
	Pregnancy (sequencer validation)	$\mathcal{D}_{3,a}$	$2 \times 45^* \begin{cases} 45 \\ 45 \end{cases}$	Kapa Hyper-Prep	Kapa dual	No	HiSeq 2000
		$\mathcal{D}_{3,b}$		Kapa Hyper-Prep	Kapa dual	No	NovaSeq
		$\mathcal{D}_{4,a}$	$2 \times 45^* \begin{cases} 45 \\ 45 \end{cases}$	Kapa Hyper-Prep	Kapa dual	No	HiSeq 2500
		$\mathcal{D}_{4,b}$		Kapa Hyper-Prep	Kapa dual	No	NovaSeq
		$\mathcal{D}_{5,a}$	$2 \times 93^* \begin{cases} 93 \\ 93 \end{cases}$	Kapa Hyper-Prep	Kapa dual	No	HiSeq 4000
		$\mathcal{D}_{5,b}$		Kapa Hyper-Prep	IDT	No	NovaSeq
	Pregnancy (chemistry validation)	$\mathcal{D}_{6,a}$	$2 \times 135^* \begin{cases} 135 \\ 135 \end{cases}$	Kapa Hyper-Prep	IDT	No	NovaSeq (V1)
		$\mathcal{D}_{6,b}$		Kapa Hyper-Prep	IDT	No	NovaSeq (V1.5)
HEMA	Hodgkin lymphoma	$\mathcal{D}_7$	179	TruSeq ChIP	-	No	HiSeq 2000/2500
	Diffuse large B-cell lymphoma	$\mathcal{D}_7$	37	TruSeq ChIP	-	No	HiSeq 2000/2500
	Multiple myeloma	$\mathcal{D}_7$	22	TruSeq ChIP	-	No	HiSeq 2000/2500
	Healthy	$\mathcal{D}_7$	242	TruSeq ChIP	-	No	HiSeq 2000/2500
	Healthy	$\mathcal{D}_8$	257	TruSeq Nano	-	No	HiSeq 2000/2500
OV	Ovarian carcinoma	$\mathcal{D}_{9,a}(\mathcal{L}_1)$	223	KAPA Hyper-Prep	IDT	No	HiSeq 4000
		$\mathcal{D}_{9,b}(\mathcal{L}_1)$	32	KAPA Hyper-Prep	-	No	HiSeq 4000
		$\mathcal{D}_{9,c}(\mathcal{L}_1)$	1	KAPA Hyper-Prep	-	No	HiSeq 2000
		$\mathcal{D}_{9,b}(\mathcal{L}_1)$	$2 \times 64^* \begin{cases} 61 \\ 2 \\ 1 \\ 64 \end{cases}$	KAPA Hyper-Prep	-	No	HiSeq 4000
		$\mathcal{D}_{9,d}(\mathcal{L}_1)$		KAPA Hyper-Prep	-	No	NovaSeq V1
		$\mathcal{D}_{9,c}(\mathcal{L}_1)$		KAPA Hyper-Prep	-	No	HiSeq 2000
	Ovarian carcinoma	$\mathcal{D}_{10}(\mathcal{L}_2)$		KAPA DNA lib. prep.	-	No	HiSeq 2500
			156	KAPA DNA lib. prep.	-	No	HiSeq 2500
	Healthy	$\mathcal{D}_{9,a}(\mathcal{L}_1)$	79	KAPA Hyper-Prep	IDT	No	HiSeq 4000
	Healthy	$\mathcal{D}_{10}(\mathcal{L}_2)$	39	KAPA DNA lib. prep.	-	No	HiSeq 2500

Table 1 (continued)

Data set	Condition/ setting	Domain	Size	Library preparation kit	Index type	Paired-end?	Sequencer
FRAG	Breast cancer	$\mathcal{D}_{11}$	51	NEBNext Enzymatic Methyl-seq	-	Yes	NovaSeq 6000
	Healthy	$\mathcal{D}_{11}$	74	NEBNext Enzymatic Methyl-seq	-	Yes	NovaSeq 6000
	Healthy	$\mathcal{D}_{12}$	57	KAPA Hyper-Prep	IDT	Yes	NovaSeq 6000

Samples in sets marked with a "∗" have been processed twice, allowing quantitative assessment of the different biases caused by the changes of sequencing protocols. Index/adaptors marked with "-" are part of the indicated library preparation kit. Each wet-lab protocol is denoted by a distinct $\mathcal{D}$ symbol

Optimal transport identifies related samples despite differences in preanalytical variables

The motivation for building our bias correction on optimal transport (OT) theory was based on its natural ability to learn a mapping between domains (i.e., sequencing protocols), called barycentric mapping. To demonstrate the large potential of OT for alleviating technical biases, we employed technical replicates produced with different sequencing protocols and investigated whether the replicate pairs could be re-identified. In Fig. 2A, we show the pairwise Bray-Curtis distances between coverage profiles derived from 66 samples prepared with both the TruSeq Nano and Kapa HyperPrep kits,

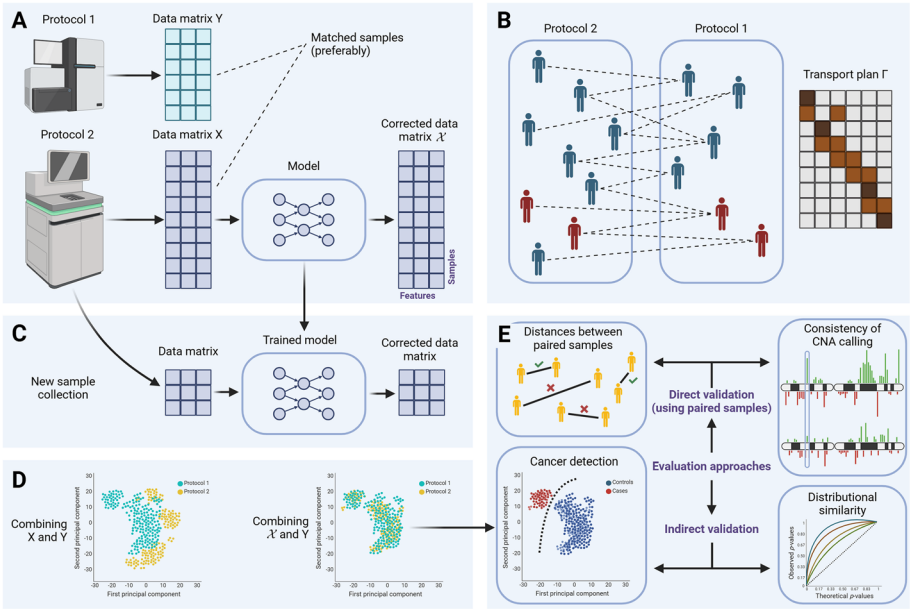

Fig. 1 Illustrative summary of our methods. **A** Given two groups of cfDNA samples differing by the sequencing pipelines they have been through, a neural network model is trained to correct the samples from protocol 2 towards protocol 1, by matching the distribution of the two groups. **B** Inference of the model is guided by the solution to the optimal transport problem, defined by the pairwise distances between samples. The solution, called transport plan, assigns samples from protocol 2 to similar samples from protocol 1. **C** Optionally, the model can be used after inference for correcting independent samples processed with protocol 2, without the need for matched samples processed with protocol 1. **D** Bias correction enables joint analysis of the two groups through better superimposition of the data distributions. **E** Summary of the validation procedures used to assess the reliability of bias correction

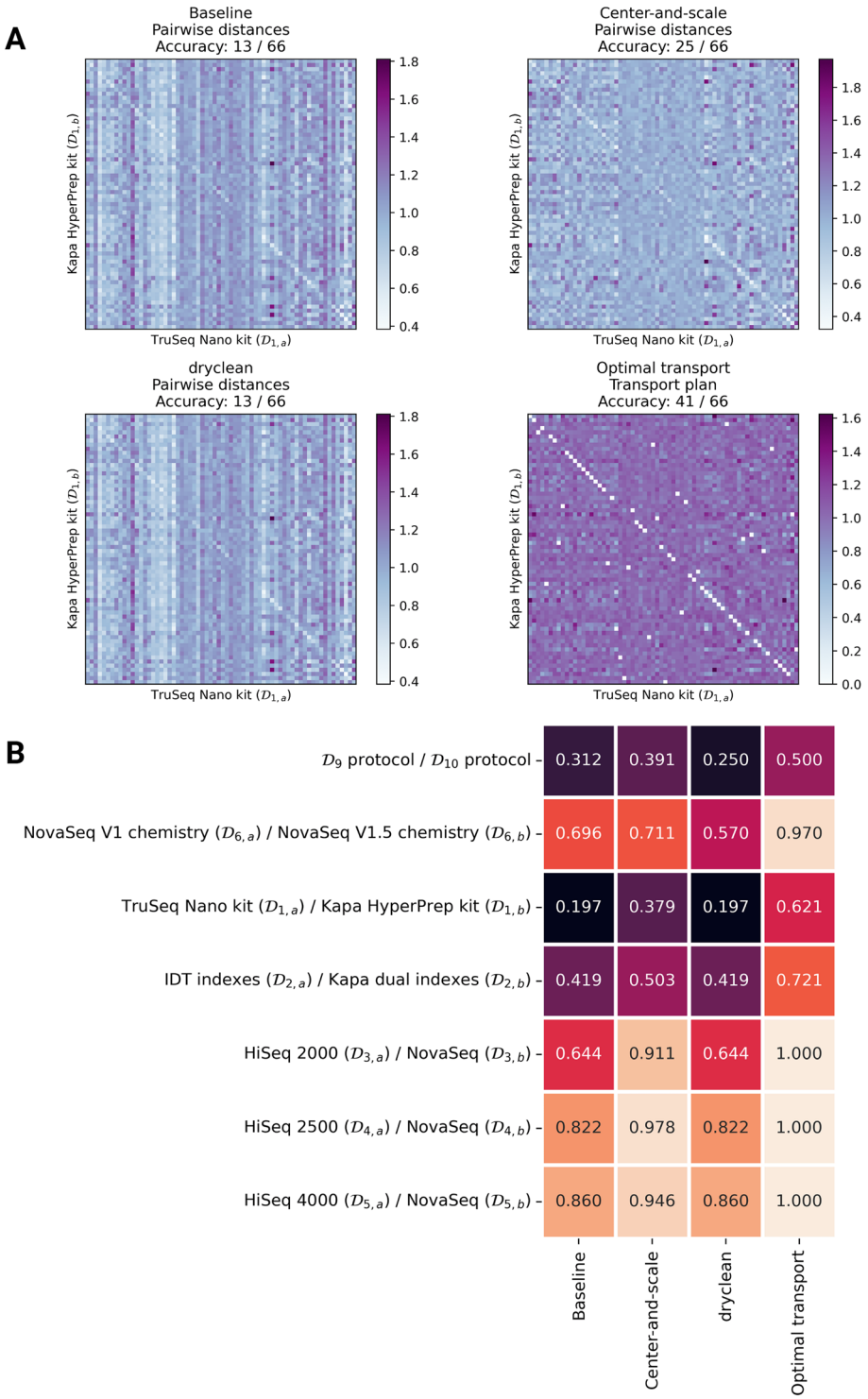

Fig. 2 Identification of paired samples using distance matrices and transport plan. **A** Pairwise distances in the absence of bias correction (top left), after center-and-scale correction (top right), after `dryclean` normalization (bottom left), and after OT-based barycentric mapping (bottom right). A pair (x, y) was considered as correctly identified when the closest sample to x from the second protocol was y and the closest sample to y from the first protocol was x. **B** Accuracy for each pair of protocols, for which replicates were available (see Table 1)

as well as after performing center-and-scale correction, `dryclean` normalization, and OT-based barycentric mapping.

Distance-based pairing ("Baseline") correctly identified 13 pairs out of 66 (19.7% accuracy). On the other hand, the use of OT improved pairing accuracy up to 62.1% (41/66). We reported heatmaps for other protocols in Additional file 1: Figs. S1–S7. A major improvement in accuracy was observed for every setting, with values ranging from 50% to 100%. In conclusion, OT has the ability to bypass technical biases when the two groups being compared are perfectly matched.

However, since OT alone cannot learn any explicit mapping between domains, it is as such incapable of processing new unseen samples. Therefore, additional algorithmic developments where needed to adapt OT to real-life settings (see Methods section). In the next sections, we compared our full method, DAGIP, with existing algorithms on various problems.

Technical biases can be accurately estimated and corrected in new unseen samples

While our method is capable of superimposing cohorts and learning a mapping between their respective sequencing protocols, there is no apriori guarantee that new unmatched profiles will be corrected in the right direction. To investigate this, we again considered the paired samples from Optimal transport identifies related samples despite differences in preanalytical variables section. Since the problem can be phrased as a typical supervised regression problem, we performed a regular 5-fold cross-validation, where the model

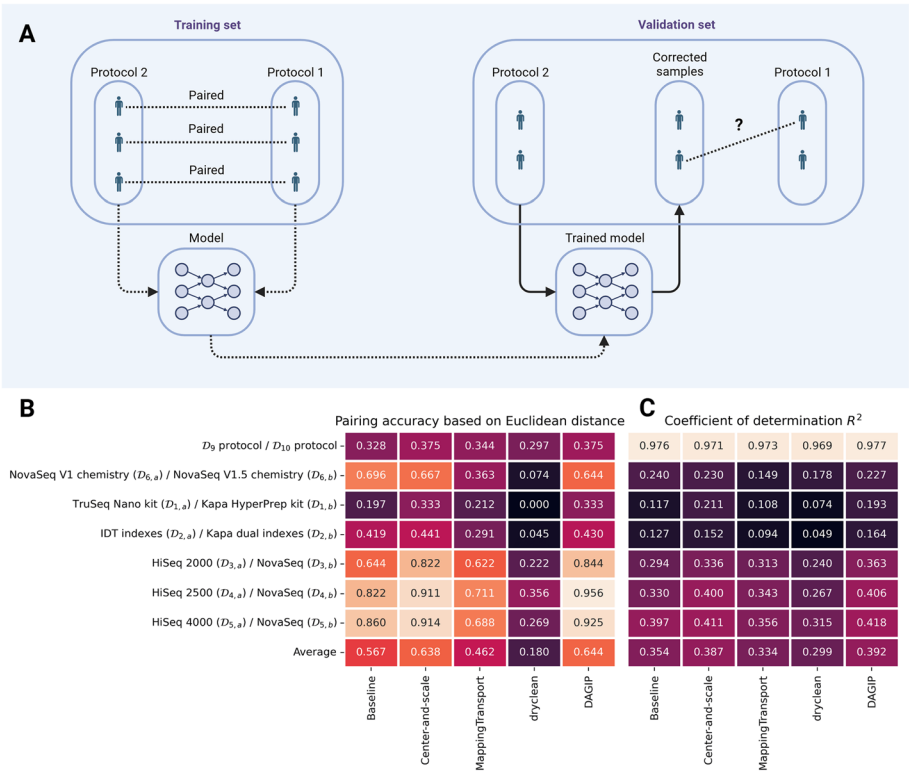

Fig. 3 Direct evaluation of bias correction approaches on coverage profiles using paired samples. **A** Illustration of the *k*-fold cross-validation procedure. **B** Sample pairing accuracy based on Bray-Curtis distance. **C** Coefficient of determination R^2 used to measure proximity between paired samples across protocols. R^2 values lie in the $[-\infty, 1]$ range

was built on 80% of the pairs and used to correct the 20% remaining samples from protocol 2 towards protocol 1, as depicted in Fig. 3A. To quantify the pairing accuracy after correction, we enumerated the pairs in which both profiles had the lowest distance across domains. Pairing accuracy has been reported for each method and protocol pair in Fig. 3B.

Pairing accuracy was the lowest when considering differences in library preparation kit (0.197 for the baseline). DAGIP systematically improved over the baseline and outperformed other methods in most settings. On average, DAGIP produced the best pairing accuracy (0.644), followed by center-and-scale (0.638), baseline (0.567), MappingTransport (0.462), and `dryclean` (0.180). The observed improvement of DAGIP turned out to be significant for all alternative methods except center-and-scale ($p > 0.005$). Significance has been reported for each method and setting in Additional file 1: Fig. S8. Pairing was done based on closest Bray-Curtis distance between samples. Indeed, Bray-Curtis distance drastically improved over all other distance metrics (e.g., Euclidean, Manhattan), as shown in Additional file 1: Fig. S9. Also, dimensionality reduction through PCA produced a major improvement in accuracy, as shown by the overall poor results on the whole coverage profiles (Additional file 1: Fig. S10).

In Fig. 3C, we also reported the proximity between coverage profiles across protocols, as measured by the coefficient of determination R^2. Performance evaluation (distance, accuracy, and R^2 computations) was based on the first principal components that contribute to 95% of the total variance, instead of the whole profiles. R^2 was relatively high for the $\mathcal{L}_1/\mathcal{L}_2$ labs (ranging from 0.958 to 0.972), as the large CNAs exhibited by by few late-stage ovarian carcinoma cases contributed to most of the variance. For the remaining protocols, results followed similar trend as for the pairing accuracy. On average, DAGIP produced the largest R^2 coefficient (0.392), followed by center-and-scale (0.387), baseline (0.354), MappingTransport (0.334), and `dryclean` (0.299). The improvement of DAGIP proved to be significant over each contender ($p < 0.005$), as reported in Additional file 1: Fig. S8. Overall, performance remained quite low, which could be indicative of the difficulty of the problem and reminiscent of the capture randomness in shallow whole-genome sequencing. Indeed, the maximum achievable R^2 score is theoretically bounded by a value that depends on the variance of technical replicates [70, 71].

Technical biases can be accurately removed from coverage and fragmentomic profiles

We next investigated whether DAGIP can help in superimposing data sets produced by different library preparation methods. We started our analyses with the FRAG data set, which contains healthy controls for which the libraries have been prepared with drastically different kits: the KAPA HyperPrep and NEBNext Enzymatic Methyl-seq kits. The rationale for adding enzymatically converted libraries to the present study was not to perform methylation calling perse, but rather including samples with distinct fragmentomic patterns. In fact, while the KAPA HyperPrep kit clearly produced a non-random end motif frequency profile, with 11 of the CCNN end motifs being among the most prominent ones (Fig. 4A), end motif frequencies from NEBNext Enzymatic Methyl-seq showed a slightly more uniform distribution, which aligns with the overall decreased mapping quality due to `bwa-meth`'s reference being a 3-letter genome [72] (Fig. 4B). NEBNext Enzymatic Methyl-seq kit also showed significantly increased proportion of short fragments (Fig. 4C). The observed trend was more intricate for larger

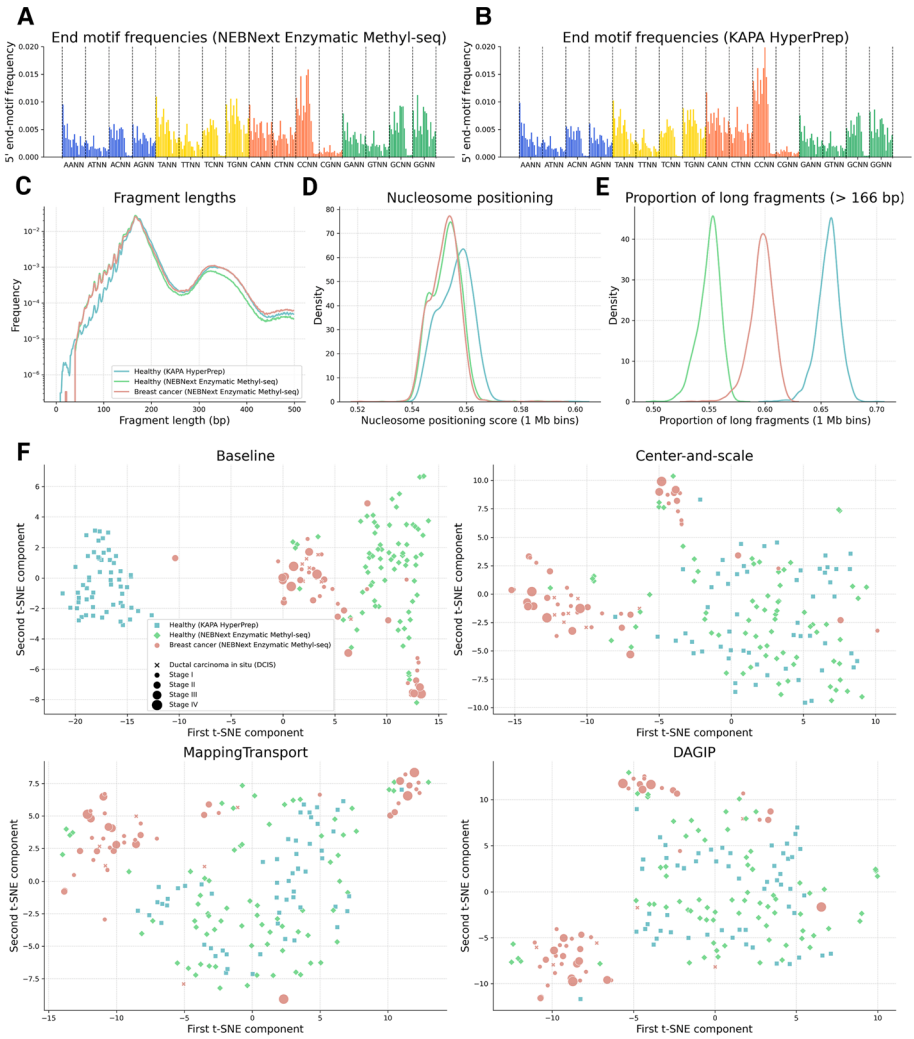

Fig. 4 Fragmentomic analysis and t-SNE visualization of the FRAG data set. **A** 4-mer end motif frequencies of control samples from protocol $\mathcal{D}_{12}$ (KAPA HyperPrep kit, shown on the right panel). **B** End motif frequencies of control samples from protocol $\mathcal{D}_{11}$ (NEBNext Enzymatic Methyl-seq kit). **C** Fragment length frequencies, computed from the $\mathcal{D}_{11}$ breast cancer cohort, the $\mathcal{D}_{11}$ control group, and the $\mathcal{D}_{12}$ control group. *y*-axis is shown in log-scale. **C** Distribution of per-bin nucleosome positioning scores. For each 1 Mb bin, the nucleosome positioning scores have been averaged across all fragments having at least one of their ends falling in that bin. Densities were averaged across each of the three groups of samples. **E** Distributions of per-bin proportions of long fragments (>166 bp). Densities were averaged across each of the three groups of samples. **F** t-SNE visualization. Marker size is proportional to the cancer stage

fragments (>300 bp), as the proportion was decreased for enzymatically converted controls compared to KAPA HyperPrep, but increased for breast cancer cases, which could be explained by non-apoptotic DNA release mechanisms and altered nuclease activity. More strikingly, nucleosome positioning scores were strongly decreased for NEBNext Enzymatic Methyl-seq, reflecting the lower mapping quality and therefore the lower consistency between these samples and the reference nucleosome map (Fig. 4D). Surprisingly, observed long (>166 bp) fragment ratios were observed as increased in breast cancer (Fig. 4E). While short (<151 bp) and longer (>220 bp) fragments are both reported as being more prominent in cancer [13, 29], a significant portion of these short

fragments have been filtered out due to their low mapping quality. In Fig. 4F, we show that these differences in fragmentomic patterns lead to the two control groups being perfectly segregated in t-SNE plots ("Baseline" panel). On the other hand, all three bias correction approaches (center-and-scale, MappingTransport, DAGIP) proved able to superimpose the control groups without any apparent discrepancy. For the HEMA and OV data sets, t-SNE plots based on coverage profiles are provided in Additional file 1: Figs. S11 and S12, respectively.

To numerically quantify the quality of bias correction on the FRAG, HEMA, and OV data sets, we first performed statistical testing between the two control groups after correction (see Methods section). In Fig. 5A, the resulting p values were compared against theoretical p values using Q-Q plots. The estimation procedure for theoretical p values is described in Indirect evaluation methods section. Theoretical p values reflect the desired situation (null hypothesis), namely the absence of bias between the two control groups. In the FRAG data set, end motif frequencies, fragment size profiles, and long fragment ratio profiles all resulted in mean absolute errors close to the 0.5 limit (Fig. 5B) and produced Q-Q curves above the diagonal (Fig. 5A), showing the strong discrepancy between the KAPA and NEBNext control groups. On the other hand, the center-and-scale approach systematically produced curves above the diagonal, which is reminiscent

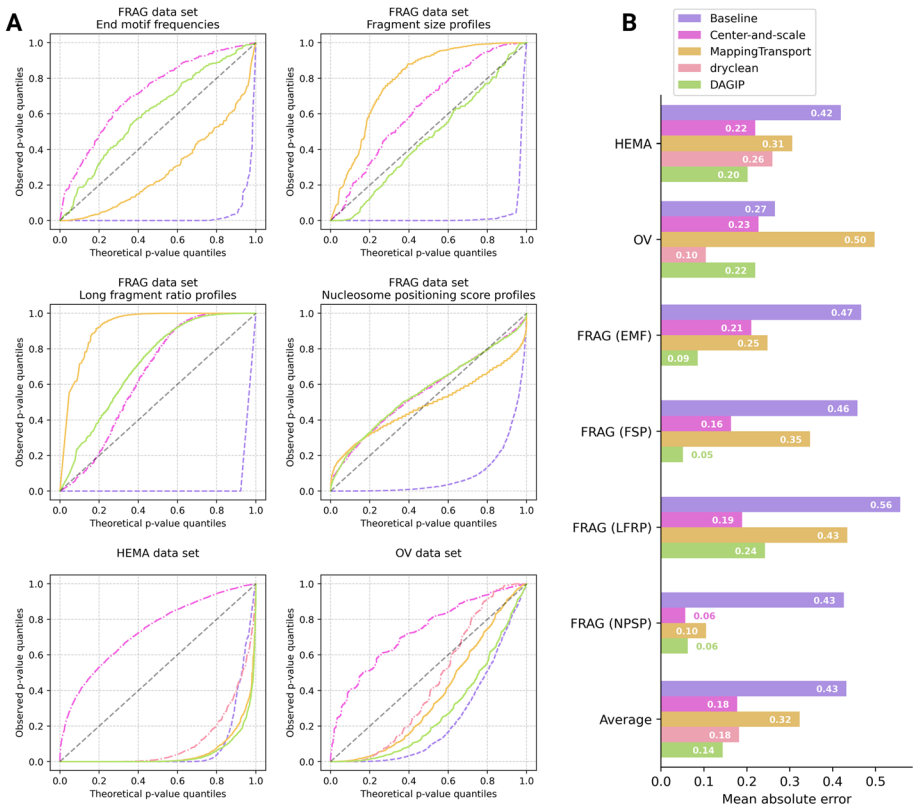

Fig. 5 Correction of the healthy samples from the OV, HEMA, and FRAG data sets. **A** Q-Q plot of p values comparing two control groups after correction. Each "empirical" p value was derived from a two-sample (two-sided) Kolmogorov-Smirnov test on a particular feature (e.g., 1 Mb bin, end motif). The background distribution of "theoretical" p values was estimated by permutation (see Methods section). **B** Mean absolute error between theoretical and observed p values. Results have not been reported for `dryclean` on the FRAG data set, as this correction method is only applicable to coverage profiles. EMF, end motif frequencies; FSP, fragment size profiles; LFRP, long fragment ratio profiles; NPSP, nucleosome positioning score profiles

of overfitting issues. While our approach DAGIP often over-corrected the samples as well, it produced the lowest mean absolute error on average (0.144), followed by center-and-scale (0.178), `dryclean` (0.182), MappingTransport (0.323), and finally the baseline (0.432). In conclusion, our method is less likely to under- or over-correct the data on average compared to other methods.

DAGIP preserves copy number aberrations across protocols

As we are fully aware of the overfitting risks associated with any domain adaptation (DA) approach, we evaluated whether the corrected CNA profiles were consistent with the original data by verifying whether somatic CNAs were conserved. For this purpose, CNAs were called from ovarian carcinoma cases (OV data set) using the `ichorCNA v0.2.0` R package (details in Additional file 1: Section 1), and their consistency across different settings was assessed. As shown in Table 1, the two labs contributing to the OV data set (denoted by $\mathcal{L}_1$ and $\mathcal{L}_2$) have 64 shared samples, which could again be used for validation. Instead of repeating the 5-fold cross-validation performed in Technical biases can be accurately estimated and corrected in new unseen samples section, we first built each of the bias correction model on the $\mathcal{L}_1$ controls and $\mathcal{L}_2$ ovarian carcinoma cases excluding the paired samples, to mimic the real-life scenario where paired samples are not necessarily available for inference. Next, each model was applied on the 64 paired ovarian carcinoma samples, and `ichorCNA` was run under various circumstances. In a first stage, we assessed the conservation of information carried in cancer cases by comparing the runs (1) and (2) illustrated in Fig. 6A. We refer to this evaluation method as intra-domain consistency assessment. However, this evaluation alone is not sufficient to ensure that CNA calling is compatible with protocol $\mathcal{L}_1$. Therefore, we also compared the CNA detection results between runs (3) and (4) (Fig. 6A) to evaluate whether the 64 cancer cases had similar copy number profiles in both domains. We refer to this evaluation approach as cross-domain consistency assessment. Results for both evaluation approaches were reported in Figs. 6 and 7.

In Fig. 6B–F, we represented copy number congruity using circular heatmaps, where the inner and outer sections correspond to intra-domain and cross-domain consistencies, respectively. The baseline had perfect intra-domain congruity, as can be observed from the entirely white inner area, and as can be expected given the absence of data correction. All bias correction methods, excluding DAGIP, had at least 20% of inconsistencies in their copy number profiles. `dryclean` caused the most changes, with 46,3% of inconsistencies. These disruptions are shown in violet (increase in copy number) and green (decrease). Interestingly, they often extend to chromosome arms or even entire chromosomes, which is directly explained by detected CNAs being of large size, and suggests that `ichorCNA` privileges changes in copy numbers over changes in segmentation. All methods, including the baseline, produced at least 30% of cross-domain inconsistencies, as shown in red (increase in copy number) and blue (decrease). Samples with lower estimated tumor fractions did not necessarily produce more mismatches, as shown by the low correlation between accuracy and tumor fraction in Fig. 6G (0.206 Pearson correlation; $p = 0.101$). When running `ichorCNA` without panel of normals (in a reference-free fashion), correlation remained non-significant (0.209 Pearson correlation; $p = 0.097$). Finally, we report strong cross-domain consistency in tumor fraction estimation

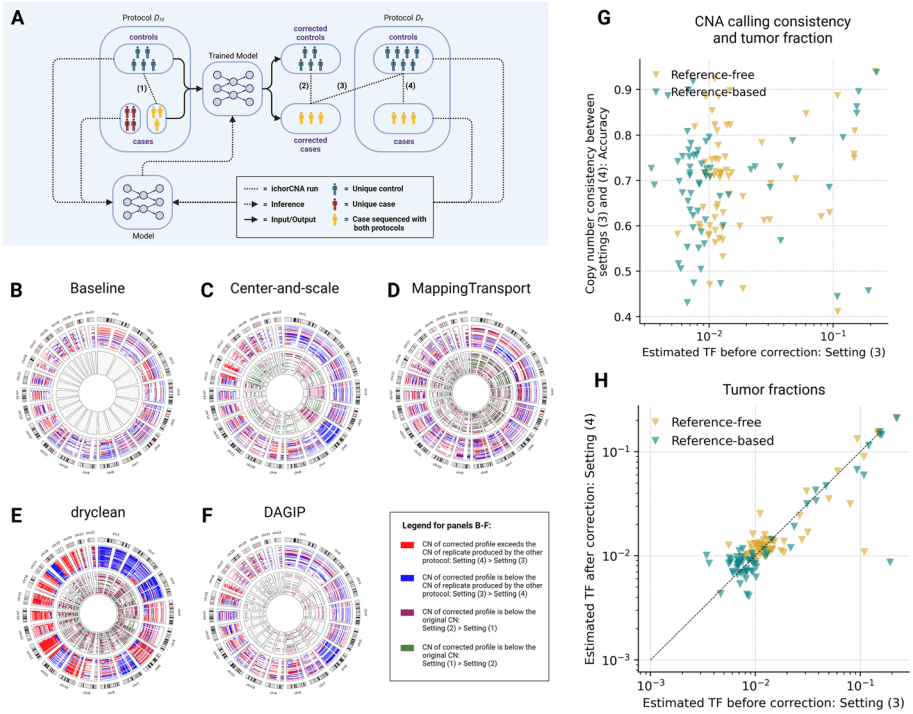

Fig. 6 Consistency of CNA calling across protocols for 64 pairs of ovarian carcinoma coverage profiles. **A** Illustration of the 4 settings in which `ichorCNA` was successively run. First validation approach, which we term intra-domain consistency, consisted in comparing CNAs between settings (1) and (2). Second approach, coined cross-domain consistency, was based on the comparison between settings (3) and (4) instead. **B**–**F** Differences in estimated copy number profiles between settings (1) and (2) (inner section), and differences between settings (3) and (4) (outer section). Red/violet corresponds to an increase in estimated copy number after correction, while blue/green corresponds to a decrease. Color white corresponds to an exact copy number match. The meaning of white is the consistency of copy numbers. Comparison was made on the estimated copy numbers (discrete values); therefore, no thresholding was required. **G** Relationship between tumor fraction estimated by `ichorCNA` (TF) and copy number cross-domain consistency for DAGIP. **H** Cross-domain consistency in tumor fraction estimation (DAGIP)

(0.876 Pearson correlation, $p = 2.83e{-}21$), as shown in Fig. 6H. When removing the panel of normals, a significance gain could be observed (0.933 Pearson correlation, $p = 3.13e{-}29$), suggesting that the use of a control group introduces some variability in the present setting.

Apart from copy number profiles, we compared the other parameters estimated by `ichorCNA`, including the log-ratios used as input to its hidden Markov model, the estimated tumor fractions, tumor ploidies, cellular prevalences, and proportions of subclonal CNAs. Additionally, we computed the segment overlap score SOV_REFINE [73] to quantify the matching of segments in copy number profiles. Both intra-domain and cross-domain results have been compiled at the top of Fig. 7. DAGIP was compared with contenders using a one-sided t-tests and p values were reported at the bottom of Fig. 7. As expected, the baseline had perfect intra-domain consistency due to the absence of data alteration (top of Fig. 7). DAGIP ranked second with an average score of 0.755, followed by center-and-scale (0.547), MappingTransport (−1.918), and `dryclean` (−3.502). Improvements of copy number profiles were significant for DAGIP over center-and-scale, MappingTransport, and `dryclean` ($p < 0.001$). This applies also

Fig. 7 Quantitative assessment of CNA calling consistency on the paired samples from the OV data set. (Top) Intra-domain and cross-domains consistency assessment of the `ichorCNA` results and parameters inferred on the 64 paired samples from ovarian carcinoma cases. (Bottom) Significance of DAGIP's improvement over each method measured by *t*-test *p* values

to the raw log-ratio profiles ($p \leq 0.002$) and tumor ploidy estimates ($p < 0.001$). In terms of cross-domain consistency, DAGIP improves over all methods (0.273), followed by center-and-scale (0.166), the baseline (0.084), MappingTransport (−2.190), and `dryclean` (−5.095). However, these performance gains were not significant for any of the `ichorCNA` results or parameters, except for `dryclean`, where DAGIP outperforms it in terms of copy number and log-ratio profiles consistencies, SOV_REFINE scores, and tumor ploidy consistencies. Finally, we added the figures generated by `ichorCNA` in Additional file: Figs. S13–S17 for 5 different cases.

DAGIP disentangles cancer signals from non-biological sources of variation

We further tested the applicability of our method to the detection of hematological cancer and investigated whether data correction preserves the signals of interest (e.g., somatic CNAs, fragmentation patterns). For this purpose, we trained simple machine learning models on the HEMA, OV, and FRAG data sets using the `scikit-learn` [74] Python library. For the HEMA data set, the samples whose libraries have been prepared

with the TruSeq ChIP Library Preparation Kit (Illumina), including 242 controls and 238 hematological cases, correspond to protocol $\mathcal{D}_7$ in Fig. 8A, while the 257 TruSeq Nano samples correspond to protocol $\mathcal{D}_8$. No cancer case was available from protocol $\mathcal{D}_8$. The FRAG data set was used similarly, as the libraries prepared with the NEBNext Enzymatic Methyl-Seq kit (51 breast cancer cases and 74 female controls) and the KAPA HyperPrep kits (57 female controls) correspond to protocols $\mathcal{D}_{11}$ and $\mathcal{D}_{12}$ in the illustration, respectively. For the OV data set, where cases and controls are available from both labels, validation was performed as illustrated in Fig. 8B instead. More details about the validation scheme can be found in Indirect evaluation methods section. Hematological cancer and ovarian carcinoma detection was based on the (corrected) coverage profiles,

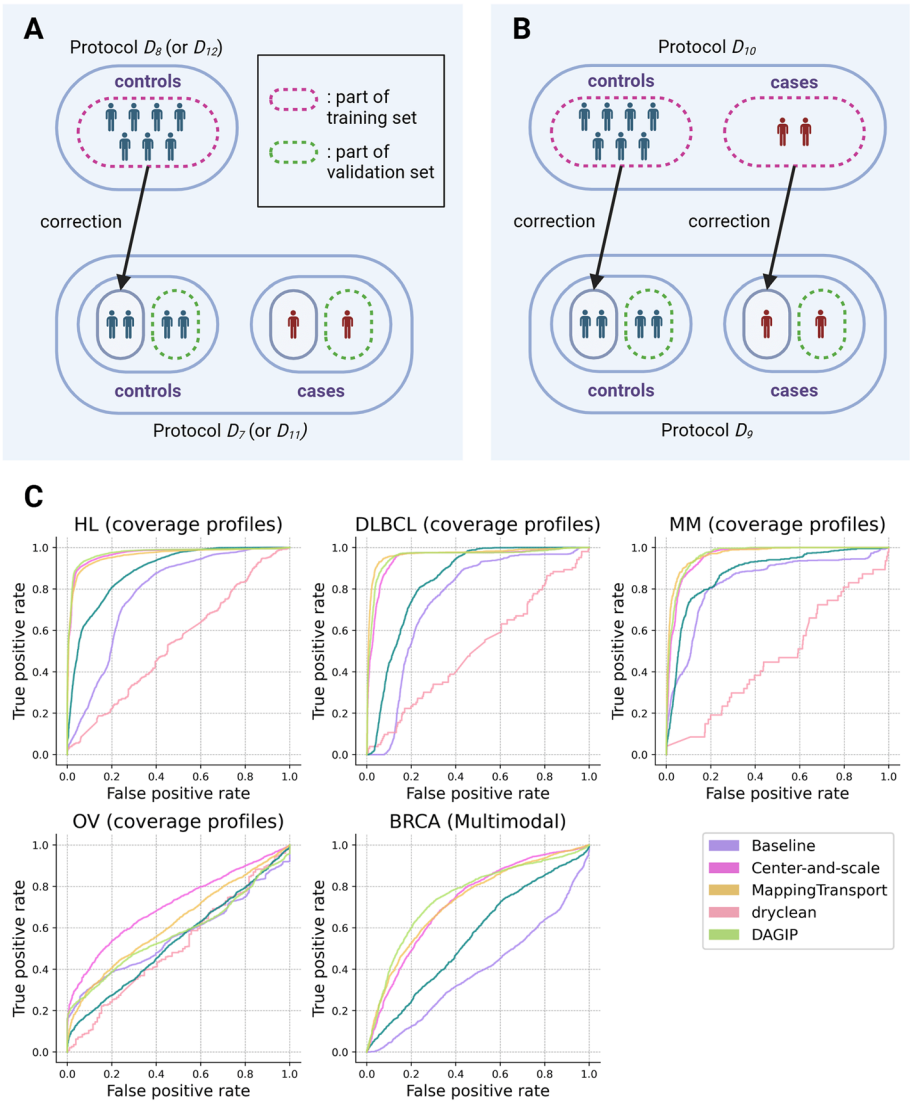

Fig. 8 Cancer detection performance assessment. **A** Depiction of the cross-validation scheme used on the HEMA and FRAG data sets. **B** Illustration of the cross-validation scheme used on the OV data set. **C** Receiver operating characteristic curve of each correction method on each data set, using a non-linear support vector machine as cancer detector. "Baseline" refers to the cancer detection performance without prior domain adaptation. `dryclean` was not applied on BRCA as the method was specifically designed for coverage profiles. HL, Hodgkin lymphoma; DLBCL, diffuse large B-cell lymphoma; MM, multiple myeloma; OV, ovarian carcinoma; BRCA, breast cancer

without any dimensionality reduction. However, each 1 Mb bin was first centered and scaled using the median and inter-quartile range (IQR), as normalization is required by most machine learning models. This preprocessing step was only performed for the computational experiment presented in this section. For the FRAG data set, breast cancer detection was based on the combination of all four fragmentomic modalities considered in this work, namely fragment size profiles, end motif frequencies, nucleosome positioning score profiles, and long fragment ratio profiles. The 4 data matrices were simply combined into a larger one, which was fed as input to the machine learning model. This approach is referred to as "multimodal" in Fig. 8C. The 64 paired samples from ovarian carcinoma cases were ensured to not be split across training and validation sets to avoid any contamination. Similarly, sequencing batches were treated similarly, to avoid batch effects from contaminating the validation set. More details about the validation scheme can be found in Indirect evaluation methods section. Hematological cancer and ovarian carcinoma detection was based on the (corrected) coverage profiles, without any dimensionality reduction.

In Fig. 8C, we show the ROC curves of each bias correction method on each use case (i.e., cancer type). Most methods improved over the baseline for hematological cancer and breast cancer detection, therefore highlighting the importance of matching data distributions when training a machine learning model on a mix of samples derived from different protocols. However, `dryclean` systematically led to random performance using a support vector machine (~0.5 AUROC), and rather low average AUROC and MCC scores as shown in Additional file 1: Fig. S18A–C, suggesting that `dryclean` is actually filtering out more information than just technical noise. All methods performed poorly on the OV data set. When averaging across the 5 cancer types, center-and-scale produced the highest AUROC (0.862), followed by DAGIP (0.854), MappingTransport (0.846), KMM (0.703), the baseline (0.659), and `dryclean` (0.589). We also reported the MCC for each data set in Additional file 1: Fig. S18C, as this metric is robust against control/case ratio imbalances. While MappingTransport produced AUROC scores similar to center-and-scale and DAGIP, the method performed comparatively worse in terms of MCC, which is reminiscent of a suboptimal supervised model calibration (the optimal prediction cutoff is not centered around 0.5). This hypothesis is also supported by the overfitting previously observed in Fig. 5A. On average, DAGIP produced the highest MCC (0.473), followed by center-and-scale (0.460), MappingTransport (0.366), KMM (0.207), the baseline (0.141), and finally `dryclean` (0.086). While the MCC confidence intervals of center-and-scale and DAGIP overlapped, the improvement of DAGIP over center-and-scale across all pathologies was significant ($p = 1.2e{-}4$, Additional file 1: Fig. S18D). Sensitivity, specificity, MCC, AUROC, and AUPR have been reported for each bias correction method, pathology, and machine learning model in Additional file 1: Tables S1–S3.

Discussion

Summary of the results

In Fig. 9, we summarized the results presented throughout the paper. Among the different methods present in the benchmark, our method DAGIP produced the best

	Direct evaluation methods				Indirect evaluation methods				
	CNA calling consistency		Pairing of replicates		p-value distributions	Cancer detection		Computations	
	Intra-domain	Cross-domain	Pairing accuracy	R^2	Mean abs. error	AUROC	MCC	Peak memory	Time
Baseline	**1.000**	0.084	0.567	0.354	0.432	0.659	0.141	**12 Mb**	**0 s**
Center-and-scale	0.547	0.166	0.638	0.387	0.178	**0.862**	0.460	23 Mb	1 s
MappingTransport	-1.918	-2.190	0.462	0.334	0.323	0.846	0.366	269 Mb	3 s
Kernel mean matching	-	-	-	-	-	0.703	0.207	13 Mb	1 s
dryclean	-3.502	-5.095	0.180	0.299	0.182	0.589	0.086	21 Mb	3737 s
DAGIP	0.755	**0.273**	**0.644**	**0.530**	**0.144**	0.854	**0.473**	56 Mb	492 s

Fig. 9 Summary of the results. By design, kernel mean matching could only be applied to the cancer detection problem. Performance metrics have been averaged across data sets. R^2, coefficient of determination; AUROC, area under the receiver operating characteristic curve; MCC, Matthews correlation coefficient

results according to most of the evaluation metrics. The baseline method produced maximal intra-domain CNA calling consistency on the OV data set as expected due to the absence of domain adaptation (DA), and center-and-scale produced the best cancer detection models according to AUROC. In fact, center-and-scale ranked second for many of the evaluation metrics, which underlines the reliability of simple approaches compared to more sophisticated ones. While `dryclean` indeed filtered some of the technical sources of variation as reflected for example by the increase in accuracy for the identification of paired samples across protocols (baseline: 0.223; `dryclean`: 0.282) or the increase in proximity of the latter measured by R^2 (baseline: 0.455; `dryclean`: 0.468), it appears that much of the biological signal has been discarded as well (AUROC: 0.589; MCC: 0.086).

Beyond pure bias removal performance, average resource consumption has been added to Fig. 9 as well. Most methods had relatively low computational cost, except MappingTransport which had higher memory requirements (269 Mb on average) and `dryclean` which was orders of magnitude slower (3737 s per run on average, $\sim 5\times$ more than DAGIP). Both MappingTransport and DAGIP rely on the computation of a transport plan, which has a complexity that is quadratic in the number of available samples. Therefore, these two methods are less scalable than other approaches and may encounter considerably longer running times on large data sets (e.g., 10,000 samples from each protocol). Regardless, all methods are scalable with respect to the number of features (e.g., fragment length frequencies), meaning that doubling their number will only double computational costs.

Study limitations and clinical validation

A major limitation of proposed solution is reminiscent of DA in general: there is no guarantee that the function learned by the bias correction algorithm accurately reflects the mechanisms of PCR amplification or sequencing. Indeed, the superimposition of the sample groups enforced by the method may be purely artificial when the two groups are unrelated. On the other hand, we devised several protection mechanisms to mitigate over-correction when no meaningful bias signal can be found in the data. Nonetheless, it is crucial that practitioners are aware of any confounder that could introduce biological discrepancy between the two groups being supplied to the tool. When cancer

cases are used to train DAGIP, they should ideally be of the same sub-type. Following the same idea, control samples should preferably be matched for age, biological sex, or any other relevant covariate. Ideally, the *same* biological samples should be sequenced under both protocols to discard any biological confounder. However, our experiments with `ichorCNA` showed that these requirements are not compulsory for improving consistency over the baseline. Similarly, the performance gains observed for supervised cancer detection compared to the baseline show that DAGIP can still converge to a meaningful solution without perfect matching of the controls.

Among the four different performance evaluation methods presented in our work, only the direct methods are deemed appropriate for clinical validation of DAGIP, which requires paired profiles originating from the same biological samples.

Extending the tool to other data modalities

While our computational experiments were based on coverage and fragmentomic profiles, our tool remains largely applicable to other modalities or any tabular biological data. First, our `DAGIP` Python package allows the user to define a custom manifold. For example, methylation ratios can be constrained to lie in the [0, 1] range by defining f and f^{-1} as the sigmoid and logit functions, respectively. Next, the user can define their own dissimilarity metric (e.g., Minkowski distance), which will only be used for solving the OT problem. The domains can also be grouped according to the hierarchical structure of the data set. For example, when samples share different clinical characteristics, they can be grouped according to these annotations and prevent incompatible samples from being matched during OT solving (e.g., cancer cases are only mapped to cancer cases). Finally, there are many hyper-parameters that can be tuned based on the dataset size to balance under- and overfitting risks.

Conclusions

In this study, we developed a novel bias correction algorithm for whole-genome cfDNA sequencing data, coined DAGIP, and showcased its applicability to different data modalities such as somatic CNA or fragmentomic profiles. More importantly, we demonstrated that joint analysis of samples derived from different sequencing pipelines not only remains possible, but can be enhanced by domain adaptation, an under-discussed solution to a ubiquitous problem in the field. Moreover, we proposed complementary evaluation approaches to assess the quality of bias correction and highlighted the improvements of DAGIP over existing methods. These findings open new avenues for the development of cfDNA-based cancer detection pipelines using data sets collected in different centers or using different wet-lab protocols.

Methods

Clinical data

We evaluated our method on four in-house data sets, each used for a different purpose. The peculiarities of each data set have been summarized in Table 1.

Blood samples were collected either into Streck cfDNA BCT or Roche Cell-Free DNA Collection Tubes. cfDNA was extracted using either the QIAamp Circulating Nucleic

Acid Kit (Qiagen) or the Maxwell automated protocol. Samples from the NIPT, HEMA, and OV data sets were pooled by batches of ~20 for multiplex sequencing using all lanes of Illumina flow cells. Each pool was sequenced either on the Illumina HiSeq 2000, HiSeq 2500, HiSeq 4000, or NovaSeq 6000 platform, single-end 1×36 bp, 1×50 bp, paired-end 2×50 bp, or 2×150 bp.

The first data set consists of 563 validation samples collected in the context of Non-Invasive Prenatal Testing (NIPT) [75] and processed twice each with different protocols. These paired samples are divided in 6 validation groups, each used to quantify the distributional shift introduced by the change of *one* preanalytical variable. The libraries of 66 biological samples have been prepared with either the TruSeq Nano DNA Sample Preparation Kit (Illumina) or the KAPA HyperPrep Kit (Roche) with Kapa Dual indexed adapters. One hundred seventy-nine samples have been prepared with either Integrated DNA Technologies (IDT) indexes or KAPA Dual indexed adapters. Forty-five samples have been processed either by the HiSeq 2000 or NovaSeq platform, 45 by the HiSeq 2500 or NovaSeq platform, 93 by the HiSeq 4000 or NovaSeq platform, and 135 by a NovaSeq platform with either V1 or V1.5 chemistry. In total, this results in 2×563 paired samples. Samples went through whole-genome low-pass sequencing at 0.1× coverage. We refer to this first data set as NIPT for short.

Our second data set (HEMA) focuses on hematological malignancies and is composed of 179 cases of Hodgkin lymphoma (HL), 37 cases of diffuse large B-cell lymphoma (DLBCL), and 22 cases of multiple myeloma (MM), as well as 499 controls from a previous study [76]. HL cases included 10 stage I, 145 stage II, 9 stage III, and 15 stage IV cases (mean age: 32, 55% of female). DLBCL cases included 1 stage I, 5 stage II, 7 stage III, 8 stage IV, and 16 unknown stages (mean age: 59, 60% of females). Finally, MM cases comprised 3 stage I, 7 stage II, 7 stage III, and 5 unknown stages (mean age: 67, 36% of females). The libraries of 242 out of the 499 controls (mean age: 69, 63% of females) have been prepared with the same kit as the hematological cancer cases, namely the TruSeq ChIP Library Preparation Kit (Illumina) [4, 77]. The TruSeq ChIP sample library preparation protocol was performed with the cfDNA extracted using cfDNA extraction kits, but not using chromatin immunoprecipitated DNA. The remaining 257 controls have been prepared with the TruSeq Nano kit. The majority of the samples had a 0.1–0.2× coverage. The number of mapped reads are reported in Additional file 1: Fig. S19. The TruSeq ChIP samples have been used previously in [16] for the validation of a supervised cancer detection approach at low-coverage.

We also analyzed female controls and ovarian carcinoma (OV) cases sequenced by us ($\mathcal{L}_1$) and/or by a different team ($\mathcal{L}_2$) [78]. In $\mathcal{L}_1$, 330 and 79 samples were collected from OV cases and controls, respectively. In $\mathcal{L}_2$, 220 and 39 samples were collected, respectively. Among these, 64 OV samples are technical replicates, as they were sequenced by both $\mathcal{L}_1$ and $\mathcal{L}_2$. OV samples were not derived from cancer patients with overt clinical disease, but rather the presence of a suspicious malignancy based on imaging. We refer to this third data set as OV. Protocols vary in multiple ways. As an example, all of the samples in $\mathcal{L}_1$ (see Table 1) have been processed with HiSeq 2500, while all samples from lab $\mathcal{L}_2$ have been sequenced by an instrument that differed from HiSeq 2500. Samples from $\mathcal{L}_1$ and $\mathcal{L}_2$ have been prepared with the KAPA

HyperPrep and KAPA DNA library preparation kits, respectively. Ovarian carcinoma samples from lab $\mathcal{L}_2$ have been manually extracted with the QIAamp Circulating Nucleic Acid kit. Let us note that samples from $\mathcal{L}_1$ belong to multiple domains, since all samples have not been processed with the same sequencer, but split across HiSeq 2000, HiSeq 4000, and NovaSeq V1. Despite the heterogeneity caused by the presence of multiple sequencers, we artificially grouped the samples in order to simplify the comparison between laboratories but also better reflect the heterogeneity and hierarchical bias structure expected to be encountered in real-life situations. Distributions of total sequencing depths are shown for protocol $\mathcal{D}_7$ and labs $\mathcal{L}_1$, $\mathcal{L}_2$ in Additional file 1: Fig. S19.

Finally, we included a paired-end sequencing data set (FRAG) which we used for fragmentomics analysis. $\mathcal{D}_{11}$ protocol was applied on 74 female controls (mean age: 71) and 51 breast cancer cases (mean age: 57). Breast cancer cases included 7 ductal in situ carcinomas, 22 stage I, 11 stage II, 7 stage III, and 4 stage IV cases. In $\mathcal{D}_{11}$, DNA was extracted with the Maxwell kit and processed with the NEBNext Enzymatic Methyl-seq kit (New England Biolabs, Ipswich, MA, USA). Sequencing was carried out on a NovaSeq 6000 S4 flowcell at 15× average depth and with a read size of 150 bp. In $\mathcal{D}_{12}$, libraries from 57 female controls were prepared using the KAPA HyperPrep kit with IDT adaptor ligation, and their sequencing was carried out on the NovaSeq 6000 platform, with a read size of 50 bp. Samples were downsampled in silico to a target depth of 0.16×, both to stabilize the variance across samples and to demonstrate the applicability of our methods to low-coverage settings. Downsampling was performed with no regard to the fragment size.

Data preprocessing

Single-end reads were first aligned to the reference genome hg38 using the Burrows-Wheeler aligner [79]. Then, read duplicates were removed using the `picard MarkDuplicates` command with default parameters [80] and remaining ones were recalibrated with the Genome Analysis Toolkit [81]. Reads flagged with secondary or supplementary alignment were discarded. No further filtering was employed. Only the 22 autosomes were considered for analysis. Reads were first counted in predefined bins of size 10 kb and smoothed by a running average of size 101 bins. Next, we removed the bins with mappability < 0.8 [63] or falling in blacklisted regions from an in-house curated list [77]. Ten-kilobyte bin counts were then summed into larger bins of size 1 Mb. Next, counts were normalized by dividing the whole profile by the median, to correct for the sequencing depth. Finally, We performed GC-correction on coverage profiles by dividing the normalized read counts by their Locally Weighted Scatterplot Smoothing (LOWESS) estimate [82], using 30% of the data points (bins) to predict the coverage. LOWESS [82] and LOESS [83] mostly differ by their implementation and build on the same underlying model. We used the Python package `statsmodels (v0.12.2)` [84] to implement LOWESS correction. A visual and quantitative assessment of GC bias is available for the HEMA data set in Additional file 1: Figs. S20–S21, and performance comparison of LOWESS's Python and LOESS's R implementations is provided in Additional file 1: Fig. S22.

For the FRAG data set, paired reads were first aligned to the reference using `bwa-meth` [72] and marked for duplicates using `Picard`. Only reads mapped in proper pairs were considered. Reads with `mapq` value below 60 or containing unknown bases within the 5′ 4-mer end motif were discarded. We extracted four types of fragmentomic features, which we refer to as fragment size profiles, end motif frequencies, genome-wide size profiles, and genome-wide nucleosome positioning profiles. Fragment size profiles were obtained by counting the insert sizes of read pairs in the 40–499 bp range. End motif frequencies were based on the frequency of each 4-mer (e.g., AATC) on the 5′ end of each fragment. Genome-wide size profiles were obtained by computing the proportion of long (>166 bp) fragments in each 1 Mb bin. Finally, genome-wide nucleosome positioning profiles consist in the average fragment positioning score in each 1 Mb bin, with respect to a reference nucleosome map. We used the most comprehensive nucleosome peak file (`GSE71378_CA01`) from [15] (GEO accession number: GSE71378) as reference. More details about nucleosome positioning scoring are given in Additional file 1: Section 2.

Given two groups of samples, the group requiring correction (source domain) is denoted by X, while the other is denoted by Y. Both X and Y are matrices with the same number of columns.

Alternative approaches included in the benchmark

We benchmarked our method against several approaches from different disciplines. The *baseline* approach simply corresponds to the absence of bias correction (besides GC-correction). The simplest bias correction approach was based on robust standardization, where the first group was re-centered and re-scaled to constrain it to have the same medians and inter-quartile ranges (IQR) as in the second group:

$$\mathcal{X}_{ik} = \frac{X_{ik} - \mu(X_{\cdot k})}{\sigma(X_{\cdot k})}\sigma(Y_{\cdot k}) + \mu(Y_{\cdot k}), \tag{1}$$

where $\mu(X_{\cdot k})$ is the median value of variable k and $\sigma(X_{\cdot k})$ is the IQR. The median and IQR have been used in place of the mean and standard deviation for robustness against outliers. We referred to this method as *center-and-scale* standardization throughout the manuscript. We also included *MappingTransport* [56, 57], a method also building on OT theory, but based on a drastically different mapping. We ran the `POT` [56] implementation of *MappingTransport* with default hyper-parameters, which is a linear model with regularization term `eta=0.001`. We also considered an instance-based DA approach called *kernel mean matching* (KMM) [85], which does not directly correct the data, but reweights the samples in the source domain to better match the two empirical distributions. Because this algorithm does not perform any bias correction per se and must be coupled with machine learning, we included it in the benchmark only for cancer detection, where supervised models were used. Finally, we included `dryclean` [66], a method specifically designed for disentangling biological and technical noise in coverage profiles using rPCA. Given that `dryclean` was specifically designed for coverage profiles, we did not apply it on the FRAG data set, which consists of other data modalities. Contrary to other methods, `dryclean` required correcting both source and target

domains, using a panel of controls from each domain, respectively. The algorithm was run with default hyper-parameters on the `Hg38` reference genome and using 4 cores.

Proposed method

In this section, we describe the modeling and algorithmic details behind DAGIP, the proposed bias correction method.

Optimal transport

Our goal is to minimize the statistical dissimilarity between two groups. Given the multivariate nature of the problem and the strong mathematical foundations behind optimal transport (OT), we propose to use the Wasserstein distance to quantify the discrepancy between protocols.

The general principle of OT is to match two probability distributions by transporting the probability mass of one distribution onto the other with minimal effort (least traveled distances), hence the name *optimal transport*. Since the number of available samples is finite, OT here consists in finding a discrete probabilistic mapping (called the *transport plan*) of the source data onto the target data, where the mapping of a source sample to a target sample bears some associated cost. We consider thus two data matrices $X \in \mathbb{R}_+^{n\times q}$ and $Y \in \mathbb{R}_+^{m\times q}$, as illustrated by the data matrices in Fig. 1A, where n and m are the sample sizes of each domain and q is the number of predefined bins. As samples are all assumed to be of equal importance, we choose uniform probabilistic weights $\nu_i = 1/n, \forall i$ and $\mu_j = 1/m, \forall j$ to define a probability distribution on these discrete samples. The cost of transportation is defined by a distance metric δ, where $\delta(X_{i\cdot}, Y_{j\cdot})$ is the distance between samples i and j.

The Wasserstein distance is defined, in its discrete form, by

$$
\begin{aligned}
W_p(C) &= \min_\Gamma \left(\sum_{i=1}^{n} \sum_{j=1}^{m} \delta(X_{i\cdot}, Y_{j\cdot})^p \Gamma_{ij} \right)^{1/p} \\
\text{s.t.} \quad & \sum_{j=1}^{m} \Gamma_{ij} = \nu_i, && \forall i, \\
& \sum_{i=1}^{n} \Gamma_{ij} = \mu_j, && \forall j, \\
& \Gamma_{ij} \geq 0, && \forall i, j,
\end{aligned}
\tag{2}
$$

where we chose $p = 2$, and δ to be the Euclidean distance for mathematical convenience. The matrix Γ, usually referred to as the transport plan and depicted in Fig. 1B, contains the amount of probability mass transferred from samples from the source domain to the target domain through optimal transport. In particular, Γ_{ij} is the probability mass transferred from point i in the source domain to point j in the target domain.

In order to make our approach scalable, we seek to turn this OT problem into a regression problem, where $\hat{X}$ is defined as the target value for X. This target is refered to as barycentric mapping in the OT literature, and defined as the projection of X that minimizes Wasserstein distance. When $p = 2$, and δ is the Euclidean distance, then there exists a closed-form solution for $\hat{X}$, and it is given by $\hat{X} = \frac{1}{m}\Gamma Y$. However, because the

corrected samples $\hat{X}$ lie in the convex hull of Y, their total variance is at most equal to the total variance of Y, potentially lower than the original variance of X. To account for this variance reduction, we correct the barycentric mapping as such:

$$\hat{X}_k \leftarrow \mu(\hat{X}_{\cdot k}) + \frac{\sum_{k=1}^{q} \sigma(X_{\cdot k})^2}{\sum_{k=1}^{q} \sigma(\hat{X}_{\cdot k})^2} \Big(\hat{X}_{ik} - \mu(\hat{X}_{\cdot k}) \Big), \tag{3}$$

where $\mu(\hat{X}_{\cdot k})$ and $\sigma(\hat{X}_{\cdot k})$ are respectively the median and IQR of column $\hat{X}_{\cdot k}$, and $\sigma(X_{\cdot k})$ is the IQR of $X_{\cdot k}$. Squared IQR is used in place of the variance for robustness against outliers.

Enforcing the data to lie on a user-defined manifold

Matrices X and Y are assumed to have a problem-specific structure, or lie on some matrix manifold. For coverage profiles, the manifold is the set of matrices with median-normalized and GC-corrected rows. For end motif frequencies, it is the multinomial manifold, encompassing the matrices with positive elements and having their rows summing up to one. X and Y are assumed to satisfy these problem-specific constraints before DA. We assume that a differentiable mapping f exists and can map any matrix to the desired manifold. We also assume that the reverse mapping f^{-1} exists. For instance, the logarithm can be used to approximate the inverse of the softmax operation (since $f(f^{-1}(X)) = X$ when each row of X already sums up to one). Similarly, the logit function reverses the sigmoid activation function. Mathematical details about the choice of f are provided in Additional file 1: Section 3.1.

Minimizing Wasserstein distance by tuning $\mathcal{X}$ directly could lead to severe overfitting. Instead, we guide the correction by enforcing the algorithm to explicitly learn a bias function g from the data, and adapt the samples as such:

$$\mathcal{X} = f(f^{-1}(X) + g(X)). \tag{4}$$

We implemented function g as a problem-specific neural architecture, comprising a bias vector, a multi-layer perceptron (MLP) and a sample-wise bias estimator. The MLP had 2 hidden layers with Parametric ReLU (PReLU) activation function and a LayerNorm layer before each activation. We also added a LayerNorm layer before the first fully connected layer. In summary, function g can be decomposed as such:

$$g(X_{i\cdot}) = b + \mathrm{MLP}_{\Theta}(X_{i\cdot}) + \mathrm{SWB}_c(X_{i\cdot}), \tag{5}$$

where b is the bias vector, MLP the multi-layer perceptron parameterized by Θ, and SWB the sample-wise bias estimator module parameterized by a vector c. The model was implemented with `PyTorch` [86] and trained with the Adam optimizer [87], which requires f to be a differentiable function. More details about neural network architecture, hyper-parameters, and choices of f functions can be found in Additional file 1: Section 3.

The outputs of our algorithm is a matrix $\mathcal{X}$, which we interpret as the surrogate of X in the target domain, and a trained neural network g, which can be used for correcting any new sample using Eq. 4.

Regularization functions

While the Wasserstein distance is often supplemented with a regularization term based on the entropy of Γ [88], we noticed that entropic regularization tends to reduce the variance of the adapted samples, ultimately collapsing them onto their centroid. This is not a desirable property because in actual high-dimensional and noisy data, the curse of dimensionality will naturally keep data points distant from each other. This creates an obstacle to the idea of mapping a source sample to the "closest" target samples. Therefore, we do not regularize Wasserstein distance based on entropy, nor on Laplacian regularization [58]. Instead, we propose a more conservative approach where the deviations of the samples from some reference (i.e., the median) should be preserved throughout the whole adaptation process.

The regularization function is defined as follows:

$$R(\mathcal{X}) = \frac{1}{2nq} \sum_{i=1}^{n} \sum_{k=1}^{q} \left(\frac{X_{ik} - \mu(X_{\cdot k})}{\sigma(X_{\cdot k})} - \frac{\mathcal{X}_{ik} - \mu(Y_{\cdot k})}{\sigma(Y_{\cdot k})} \right)^2, \tag{6}$$

where $\mu(X_{\cdot k})$ is the median over column k of X, and σ is the IQR. This regularization function is meant to preserve the quantiles (akin to the z-scores) across the two domains after correcting and merging them.

Additionally, we further regularized our model by adding a L2 penalty term, defined as the sum of squares of all neural network parameters contained in Θ.

Univariate Wasserstein distance

The different methodological choices made so far do not directly constrain the marginal distributions to be similar across domains. Indeed, the median of the barycentric mapping is unlikely to be exactly equal to the median of the target domain, and the same applies to the IQR. Therefore, we added another loss term to the objective function, based on univariate Wasserstein distance. Under normality assumption, this metric has a closed form, in which we again replaced the mean and standard deviation by the median and IQR:

$$u_k(\mathcal{X}_{\cdot k}, X_{\cdot k}) = (\mu(\mathcal{X}_{\cdot k}) - \mu(X_{\cdot k}))^2 + \sigma(\mathcal{X}_{\cdot k})^2 + \sigma(X_{\cdot k})^2 - 2\sigma(\mathcal{X}_{\cdot k})\sigma(X_{\cdot k}). \tag{7}$$

Objective function

The total objective function $\mathcal{L}$ is composed of the different loss and regularization terms described so far:

$$\mathcal{L}(b, c, \Theta) = \frac{1}{nq} \sum_{i=1}^{n} \sum_{k=1}^{q} \frac{1}{\sigma(Y_{\cdot k})} \left(\mathcal{X}_{ik} - \hat{X}_{ik} \right)^2 + \lambda_1 R(\mathcal{X})^2 + \lambda_2 \sum_{\theta \in \Theta} \|\theta\|_2^2 + \lambda_3 \frac{1}{q} \sum_{k=1}^{q} u_k(\mathcal{X}_{\cdot k}, X_{\cdot k}), \tag{8}$$

where the first term is the IQR-adjusted Wasserstein distance, $\|\cdot\|_2^2$ is the L2 norm, and $(\lambda_1, \lambda_2, \lambda_3)$ are hyper-parameters. This function is optimized using the Adam optimizer as described in Additional file 1: Section 3.3.

Benchmarking and performance assessment

We propose four different evaluation approaches to quantify the relevance of each bias correction method. They can be categorized as either direct or indirect methods, depending on whether or not technical replicates are used across protocols. Indeed, the availability of paired samples from two different protocols enables direct quantification of the bias separating the two domains and is crucial for clinical validation.

Most of our computational experiments required fitting each bias correction model on a subset of the data ("training set") and then applying the model on an independent subset ("validation set"). Therefore, each benchmarked method required a training phase where only a fraction of the samples were used. For center-and-scale, we simply computed the median μ and IQR σ on the training set. The implementations of KMM and MatchingTransport both offered `fit` and `transform` methods, allowing to separate the training and bias correction phases. For the sake of user-friendliness, our implementation of DAGIP implements the same methods as well. Let us note that for each aforementioned algorithm, only the samples of the source domain undergo bias correction. Because `dryclean` is not a DA method but a standardization tool, it required correction of all samples from all protocols. Therefore, we standardized each sample from a given domain using a panel of normals (i.e., controls) made of all controls available in the training set for that domain.

Direct evaluation approaches

Pairing of the replicates when both groups contain the same biological samples For the groups in which biological samples have been sequenced twice in the NIPT and OV data sets (marked with "$*$" in Table 1), we applied our correction algorithm and computed accuracy, measured as the fraction of pairs (x, y) for which y is the closest profile to x in the target domain and x is the closest profile to y in the source domain. We used the Bray-Curtis distance as defined in the `SciPy` Python package [89] to determine the proximity between samples:

$$BC = \frac{\sum_{k=1}^{q} |x_k - y_k|}{\sum_{k=1}^{q} |x_k + y_k|}. \tag{9}$$

To alleviate the underdetermination introduced by the high-dimensionality of the data, Bray-Curtis distance was measured on the first principal components (PCs) that explain 95% of the total variance. The number of selected PCs has been reported for each setting in Additional file 1: Fig. S23. The purpose of this dimensionality reduction is to remove the inherent noise present in the data. Additionally, we reported the coefficient of determination denoted by R^2, which reflects the amount of variance explained by the correction. A $R^2 = 1$ score highlights a perfect superimposition of paired samples, while a negative R^2 score reflects a situation worse than simply predicting the mean of the target domain. Significance of the R^2 differences was assessed by a z-test [90].

Consistency of copy number aberration analysis We next assessed the ability of the different methods to ensure the consistency across protocols. In particular, we performed CNA calling on the 64 paired samples from ovarian carcinoma cases (OV data set) and

assessed the consistency of the results across different settings. The four different settings were depicted in Fig. 6A.

Indirect evaluation methods

Cancer detection based on supervised learning We trained three binary classifiers, namely logistic regressions, random forests, and kernel support vector machines using the `scikit-learn` [74] Python package. Machine learning models were trained on the features described in Data preprocessing section (e.g., 1 Mb bins, end motif frequencies) directly, without any dimensionality reduction (e.g., PCA) or feature selection. We employed a problem-specific variant of the *k*-fold cross-validation scheme that we describe in Additional file 1: Section 4.2.1.

The cancer detection performance of each supervised model was quantified based on widely used metrics such as the area under the receiver operating characteristic curve (AUROC). We also added Matthews correlation coefficient (MCC) to deal with the cases/controls ratio imbalance. MCC was calculated using the default prediction cutoff (0.5), therefore penalizing machine learning models that are improperly calibrated due to distributional shifts between the training and validation sets.

To reduce fluctuations related to the random splitting of sample groups during *k*-fold as well as the randomness introduced by each bias correction method, we repeated the whole procedure 30 times and averaged the results. However, cross-validation was repeated only 3 times for `dryclean` due to excessive computation time. Finally, we reported the confidence intervals for MCC and AUROC and performed *t*-tests to assess the significance of the results. Statistical testing involving `dryclean` was based on an independent samples *t*-test instead of a paired *t*-test due to the missing 27 repeats. Results were considered significant when $p < 0.005$.

Assessing bias correction quality from p-value distributions We made *p* value Q-Q plots to detect over- or underfitting issues in the benchmarked methods, as described in Additional file 1: Section 4.2. Depending on the position of the curve relative to the diagonal in the Q-Q plots, it is possible to visualize whether the data has been over-corrected or under-corrected, indicated by observed *p* values that exceed or fall below the theoretical *p* values, respectively. Finally, to numerically quantify the quality of the correction, we calculated the mean absolute error between theoretical and observed *p* values.

Supplementary information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-025-03511-y.

Additional file 1: this file includes all supplementary figures, tables, as well as technical details about the methodology.

Additional file 2: review history.

Acknowledgements
The resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation—Flanders (FWO) and the Flemish Government. Illustrations have been created with `BioRender.com`.

Peer review information
Andrew Cosgrove was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Review history
The review history is available as Additional file 2.

Authors' contributions
J.R.V. and Y.M. supervised this project. A.P., J.R.V., and T.J. designed the experiments. A.P., Y.M., and D.R. designed the computational methods. S.T., A.V., and T.J. collected the data. A.P. performed the computational experiments. A.P., T.J., S.T., D.R., Y.M., and J.R.V. wrote the first draft of the manuscript. All authors read, revised, and approved the final manuscript.

Funding
Antoine Passemiers is funded by a Research Foundation—Flanders (FWO) doctoral fellowship (1SB2721N). Stefania Tuveri is also funded by FWO (1S74420N). Tatjana Jatsenko is funded by Agentschap Innoveren en Ondernemen (VLAIO; Flanders Innovation & Entrepreneurship grant HBC.2018.2108). Joris Robert Vermeesch is funded by FWO (G080217N, S003422N) and KU Leuven (no. C1/018). Daniele Raimondi is funded by a FWO post-doctoral fellowship (12Y5623N). Yves Moreau is funded by (1) Research Council KU Leuven: Symbiosis 4 (C14/22/125); CELSA Active Learning, (2) Innovative Medicines Initiative: MELLODY, (3) Flemish Government (ELIXIR Belgium, IWT, FWO 06260, FWO SBO MICADO S003422N, VLAIO ATHENA HBC.2019.2528), and (4) Impulsfonds AI: VR 2019 2203 DOC.0318/1QUATER Kenniscentrum Data en Maatschappij. An Coosemans is funded by the Flemish Cancer Society (2016/10728/2603). ctDNA samples within the Trans-IOTA study were financed by Kom Op Tegen Kanker (Stand Up to Cancer).

Data availability
Bed files from the hematological cancer cases and healthy controls constituting domains $\mathcal{D}_7$ (HEMA data set, see Table 1) and domain $\mathcal{D}_9$ (lab $\mathcal{L}_1$, OV data set) have been previously deposited to ArrayExpress under accession number E-MTAB-10934 [91]. Raw sequencing data from ovarian carcinoma and healthy controls (OV data set) from lab $\mathcal{L}_2$ has been previously deposited to the European Genome-phenome Archive (EGA) under dataset no. EGAD00001007748 [92] and is available under controlled access. Similarly, raw sequencing data is available for the FRAG data set at EGA (data set no. EGAD50000000257) under controlled access [93]. The remaining samples (NIPT data set) are in-house samples for which there was no explicit consent to publish the sequencing data. For these samples, only the processed data has been made available. All the processed data required to reproduce our results (coverage and fragmentomic profiles) have been uploaded to FigShare (DOI: 10.6084/m9.figshare.24459304) [94]. Our tool, as well as the scripts used to generate our results, can be found on Zenodo [95] with the URL https://doi.org/10.5281/zenodo.14503339, under the MIT license. The code is also available under the MIT license on GitHub [96] with the URL https://github.com/JorisVermeeschLab/DAGIP.

Declarations

Ethics approval and consent to participate
The data sets used in the present work have been collected during studies previously approved by the ethical committee of the University Hospitals Leuven (study protocols S/50623, S/51375, S/55904, S/56534, S/57144, S/57999, S/59207, S/62285, S/62795, S/62817, S/63240, S/64035, S/64205, S/66450, and S/67127). In accordance with the Declaration of Helsinki, written informed consent was obtained from all participants to release information for study purposes. There was no option to opt-out. This consent was approved by University Hospitals Leuven ethics committee, permitting the use of these data for research purposes. All experiments and experimental methods have been designed in compliance with the Declaration of Helsinki.

Competing interests
A.C. is a contracted researcher for Oncoinvent AS and Novocure and a consultant for Sotio AS, Epics Therapeutics SA, and Molecular Partners.

Author details
[1]Dynamical Systems, Signal Processing and Data Analytics (STADIUS), KU Leuven, Leuven, Belgium. [2]Laboratory for Cytogenetics and Genome Research, Department of Human Genetics, KU Leuven, Leuven, Belgium. [3]Department of Gynaecology and Obstetrics, University Hospitals Leuven, Leuven, Belgium. [4]Division of Gynaecological Oncology, Leuven Cancer Institute, KU Leuven, Leuven, Belgium. [5]Center for Cancer Biology, VIB, Leuven, Belgium. [6]Laboratory for Translational Genetics, Department of Human Genetics, KU Leuven, Leuven, Belgium. [7]Department of Oncology, Laboratory of Tumor Immunology and Immunotherapy, Leuven Cancer Institute, Leuven, Belgium. [8]Department of Oncology, Molecular Digestive Oncology, KU Leuven, Leuven, Belgium. [9]Department of Human Genetics, Laboratory of Genetics of Malignant Diseases, KU Leuven, Leuven, Belgium. [10]Department of Hematology, University Hospitals Leuven, Leuven, Belgium. [11]Institut de Génétique Moléculaire de Montpellier (IGMM), Université de Montpellier, Montpellier, France.

Received: 24 November 2023 Accepted: 21 February 2025
Published online: 07 March 2025

References

1. Bianchi DW, Wilkins-Haug L. Integration of noninvasive DNA testing for aneuploidy into prenatal care: what has happened since the rubber met the road? Clin Chem. 2014;60(1):78–87.
2. Leila D, Brison N, Van den Bogaert K, Dehaspe L, Janssens K, Blaumeiser B, et al. Incidence of uncommon fetal aneuploidies detected by non-invasive prenatal testing. In: 17th annual Belgian Society of Human Genetics meeting: human genetics goes somatic. Kraainem: Belgian Society of Human Genetics; 2017. p. 100.
3. De Vlaminck I, Valantine HA, Snyder TM, Strehl C, Cohen G, Luikart H, et al. Circulating cell-free DNA enables noninvasive diagnosis of heart transplant rejection. Sci Transl Med. 2014;6(241):241ra77.
4. Lenaerts L, Vandenberghe P, Brison N, Che H, Neofytou M, Verheecke M, et al. Genomewide copy number alteration screening of circulating plasma DNA: potential for the detection of incipient tumors. Ann Oncol. 2019;30(1):85–95.
5. Tan EM, Schur PH, Carr RI, Kunkel HG, et al. Deoxybonucleic acid (DNA) and antibodies to DNA in the serum of patients with systemic lupus erythematosus. J Clin Investig. 1966;45(11):1732–40.
6. Koffler D, Agnello V, Winchester R, Kunkel HG, et al. The occurrence of single-stranded DNA in the serum of patients with systemic lupus erythematosus and other diseases. J Clin Investig. 1973;52(1):198–204.
7. Jiang P, Lo YD. The long and short of circulating cell-free DNA and the ins and outs of molecular diagnostics. Trends Genet. 2016;32(6):360–71.
8. Newman AM, Bratman SV, To J, Wynne JF, Eclov NC, Modlin LA, et al. An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. Nat Med. 2014;20(5):548–54.
9. Phallen J, Sausen M, Adleff V, Leal A, Hruban C, White J, et al. Direct detection of early-stage cancers using circulating tumor DNA. Sci Transl Med. 2017;9(403):eaan2415.
10. Vandenberghe P, Wlodarska I, Tousseyn T, Dehaspe L, Dierickx D, Verheecke M, et al. Non-invasive detection of genomic imbalances in Hodgkin/Reed-Sternberg cells in early and advanced stage Hodgkin's lymphoma by sequencing of circulating cell-free DNA: a technical proof-of-principle study. Lancet Haematol. 2015;2(2):e55–65.
11. Lenaerts L, Che H, Brison N, Neofytou M, Jatsenko T, Lefrere H, et al. Breast cancer detection and treatment monitoring using a noninvasive prenatal testing platform: utility in pregnant and nonpregnant populations. Clin Chem. 2020;66(11):1414–23.
12. Cristiano S, Leal A, Phallen J, Fiksel J, Adleff V, Bruhm DC, et al. Genome-wide cell-free DNA fragmentation in patients with cancer. Nature. 2019;570(7761):385–9.
13. Mouliere F, Chandrananda D, Piskorz AM, Moore EK, Morris J, Ahlborn LB, et al. Enhanced detection of circulating tumor DNA by fragment size analysis. Sci Transl Med. 2018;10(466):eaat4921.
14. Mouliere F, Robert B, Arnau Peyrotte E, Del Rio M, Ychou M, Molina F, et al. High fragmentation characterizes tumour-derived circulating DNA. PLoS ONE. 2011;6(9):e23418.
15. Snyder MW, Kircher M, Hill AJ, Daza RM, Shendure J. Cell-free DNA comprises an in vivo nucleosome footprint that informs its tissues-of-origin. Cell. 2016;164(1–2):57–68.
16. Che H, Jatsenko T, Lenaerts L, Dehaspe L, Vancoillie L, Brison N, et al. Pan-cancer detection and typing by mining patterns in large genome-wide cell-free DNA sequencing datasets. Clin Chem. 2022;68(9):1164–76.
17. Stanley KE, Jatsenko T, Tuveri S, Sudhakaran D, Lannoo L, Van Calsteren K, et al. Cell type signatures in cell-free DNA fragmentation profiles reveal disease biology. Nat Commun. 2024;15(1):2220.
18. Lo YD, Han DS, Jiang P, Chiu RW. Epigenetics, fragmentomics, and topology of cell-free DNA in liquid biopsies. Science. 2021;372(6538):eaaw3616.
19. Jiang P, Sun K, Tong YK, Cheng SH, Cheng TH, Heung MM, et al. Preferred end coordinates and somatic variants as signatures of circulating tumor DNA associated with hepatocellular carcinoma. Proc Natl Acad Sci. 2018;115(46):E10925–33.
20. Chan KA, Jiang P, Sun K, Cheng YK, Tong YK, Cheng SH, et al. Second generation noninvasive fetal genome analysis reveals de novo mutations, single-base parental inheritance, and preferred DNA ends. Proc Natl Acad Sci. 2016;113(50):E8159–68.
21. Jiang P, Chan CW, Chan KA, Cheng SH, Wong J, Wong VWS, et al. Lengthening and shortening of plasma DNA in hepatocellular carcinoma patients. Proc Natl Acad Sci. 2015;112(11):E1317–25.
22. Curtis SD, Summers M, Cohen JD, Wang Y, Nehme N, Popoli M, et al. Identifying cancer patients from GC-patterned fragment ends of cell-free DNA. medRxiv. 2022;2022–08.
23. Jiang P, Xie T, Ding SC, Zhou Z, Cheng SH, Chan RW, et al. Detection and characterization of jagged ends of double-stranded DNA in plasma. Genome Res. 2020;30(8):1144–53.
24. Han DS, Lo YD. The nexus of cfDNA and nuclease biology. Trends Genet. 2021;37(8):758–70.
25. Han DS, Ni M, Chan RW, Wong DK, Hiraki LT, Volpi S, et al. Nuclease deficiencies alter plasma cell-free DNA methylation profiles. Genome Res. 2021;31(11):2008–21.
26. Ding SC, Chan RW, Peng W, Huang L, Zhou Z, Hu X, et al. Jagged ends on multinucleosomal cell-free DNA serve as a biomarker for nuclease activity and systemic lupus erythematosus. Clin Chem. 2022;68(7):917–26.
27. Liu MC, Oxnard G, Klein E, Swanton C, Seiden M, Liu MC, et al. Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA. Ann Oncol. 2020;31(6):745–59.
28. Bae M, Kim G, Lee TR, Ahn JM, Park H, Park SR, et al. Integrative modeling of tumor genomes and epigenomes for enhanced cancer diagnosis by cell-free DNA. Nat Commun. 2023;14(1):2017.
29. Peneder P, Stütz AM, Surdez D, Krumbholz M, Semper S, Chicard M, et al. Multimodal analysis of cell-free DNA whole-genome sequencing for pediatric cancers with low mutational burden. Nat Commun. 2021;12(1):3230.
30. Che H, Stanley K, Jatsenko T, Thienpont B, Vermeesch JR. Expanded knowledge of cell-free DNA biology: potential to broaden the clinical utility. Extracell Vesicles Circ Nucleic Acids. 2022;3:199–217.
31. Diaz LA Jr, Bardelli A. Liquid biopsies: genotyping circulating tumor DNA. J Clin Oncol. 2014;32(6):579.
32. El Messaoudi S, Rolet F, Mouliere F, Thierry AR. Circulating cell free DNA: preanalytical considerations. Clin Chim Acta. 2013;424:222–30.
33. Bronkhorst AJ, Aucamp J, Pretorius PJ. Cell-free DNA: preanalytical variables. Clin Chim Acta. 2015;450:243–53.
34. Meddeb R, Pisareva E, Thierry AR. Guidelines for the preanalytical conditions for analyzing circulating cell-free DNA. Clin Chem. 2019;65(5):623–33.

35. Till JE, Black TA, Gentile C, Abdalla A, Wang Z, Sangha HK, et al. Optimization of sources of circulating cell-free DNA variability for downstream molecular analysis. J Mol Diagn. 2021;23(11):1545–52.
36. van Dessel LF, Vitale SR, Helmijr JC, Wilting SM, van der Vlugt-Daane M, Oomen-de Hoop E, et al. High-throughput isolation of circulating tumor DNA: a comparison of automated platforms. Mol Oncol. 2019;13(2):392–402.
37. Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, Hegarty R, et al. Characterizing and measuring bias in sequence data. Genome Biol. 2013;14(5):1–20.
38. Sato MP, Ogura Y, Nakamura K, Nishida R, Gotoh Y, Hayashi M, et al. Comparison of the sequencing bias of currently available library preparation kits for Illumina sequencing of bacterial genomes and metagenomes. DNA Res. 2019;26(5):391–8.
39. Costello M, Fleharty M, Abreu J, Farjoun Y, Ferriera S, Holmes L, et al. Characterization and remediation of sample index swaps by non-redundant dual indexing on massively parallel sequencing platforms. BMC Genomics. 2018;19(1):1–10.
40. Sinha R, Stanley G, Gulati GS, Ezran C, Travaglini KJ, Wei E, et al. Index switching causes "spreading-of-signal" among multiplexed samples in Illumina HiSeq 4000 DNA sequencing. BioRxiv. 2017;125724.
41. Browne PD, Nielsen TK, Kot W, Aggerholm A, Gilbert MTP, Puetz L, et al. GC bias affects genomic and metagenomic reconstructions, underrepresenting GC-poor organisms. GigaScience. 2020;9(2):giaa008.
42. Sun B, Feng J, Saenko K. Return of frustratingly easy domain adaptation. In: Proceedings of the AAAI conference on artificial intelligence. vol. 30. Washington, DC: AAAI Press; 2016.
43. Ben-David S, Blitzer J, Crammer K, Kulesza A, Pereira F, Vaughan JW. A theory of learning from different domains. Mach Learn. 2010;79(1):151–75.
44. Mancini M, Bulo SR, Caputo B, Ricci E. Best sources forward: domain generalization through source-specific nets. In: 2018 25th IEEE international conference on image processing (ICIP). New York City: IEEE; 2018. p. 1353–7.
45. Ganin Y, Ustinova E, Ajakan H, Germain P, Larochelle H, Laviolette F, et al. Domain-adversarial training of neural networks. J Mach Learn Res. 2016;17(59):1–35.
46. Bousmalis K, Trigeorgis G, Silberman N, Krishnan D, Erhan D. Domain separation networks. Adv Neural Inf Process Syst. 2016;29:343–51.
47. Hoffman J, Tzeng E, Park T, Zhu JY, Isola P, Saenko K, et al. Cycada: Cycle-consistent adversarial domain adaptation. In: International conference on machine learning. Cambridge: Pmlr; 2018. p. 1989–98.
48. Laradji IH, Babanezhad R. M-ADDA: unsupervised domain adaptation with deep metric learning. Domain Adapt Vis Underst. 2020;17–31.
49. Cao J, Katzir O, Jiang P, Lischinski D, Cohen-Or D, Tu C, et al. Dida: disentangled synthesis for domain adaptation. 2018. arXiv preprint arXiv:1805.08019.
50. Chen L, Chen H, Wei Z, Jin X, Tan X, Jin Y, et al. Reusing the task-specific classifier as a discriminator: discriminator-free adversarial domain adaptation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. New York City: IEEE; 2022. p. 7181–90.
51. Long M, Cao Y, Cao Z, Wang J, Jordan MI. Transferable representation learning with deep adaptation networks. IEEE Trans Pattern Anal Mach Intell. 2018;41(12):3071–85.
52. Zellinger W, Moser BA, Grubinger T, Lughofer E, Natschläger T, Saminger-Platz S. Robust unsupervised domain adaptation for neural networks via moment alignment. Inf Sci. 2019;483:174–91.
53. Chen C, Fu Z, Chen Z, Jin S, Cheng Z, Jin X, et al. Homm: higher-order moment matching for unsupervised domain adaptation. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34. Washington, DC: AAAI Press; 2020. p. 3422–9.
54. Lao Q, Jiang X, Havaei M. Hypothesis disparity regularized mutual information maximization. In: Proceedings of the AAAI conference on artificial intelligence. vol. 35. Washington, DC: AAAI Press; 2021. p. 8243–51.
55. Fatras K, Sejourne T, Flamary R, Courty N. Unbalanced minibatch optimal transport; applications to domain adaptation. In: Meila M, Zhang T, editors. Proceedings of the 38th international conference on machine learning. vol. 139 of proceedings of machine learning research. Cambridge: PMLR; 2021. pp. 3186–97.
56. Flamary R, Courty N, Gramfort A, Alaya MZ, Boisbunon A, Chambon S, et al. POT: Python optimal transport. J Mach Learn Res. 2021;22(78):1–8.
57. Perrot M, Courty N, Flamary R, Habrard A. Mapping estimation for discrete optimal transport. Adv Neural Inf Process Syst. 2016;29:4197–205.
58. Flamary R, Courty N, Rakotomamonjy A, Tuia D. Optimal transport with Laplacian regularization. In: NIPS 2014, workshop on optimal transport and machine learning. Montréal: Neural Information Processing Systems Foundation; 2014.
59. Azodi CB, Tang J, Shiu SH. Opening the black box: interpretable machine learning for geneticists. Trends Genet. 2020;36(6):442–55.
60. Benjamini Y, Speed TP. Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Res. 2012;40(10):e72.
61. Alkan C, Kidd JM, Marques-Bonet T, Aksay G, Antonacci F, Hormozdiari F, et al. Personalized copy number and segmental duplication maps using next-generation sequencing. Nat Genet. 2009;41(10):1061–7.
62. Chandrananda D, Thorne NP, Ganesamoorthy D, Bruno DL, Benjamini Y, Speed TP, et al. Investigating and correcting plasma DNA sequencing coverage bias to enhance aneuploidy discovery. PLoS One. 2014;9(1):e86993.
63. Adalsteinsson VA, Ha G, Freeman SS, Choudhury AD, Stover DG, Parsons HA, et al. Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. Nat Commun. 2017;8(1):1–13.
64. Cheung MS, Down TA, Latorre I, Ahringer J. Systematic bias in high-throughput sequencing data and its correction by BEADS. Nucleic Acids Res. 2011;39(15):e103.
65. Larson NB, Larson MC, Na J, Sosa CP, Wang C, Kocher JP, et al. Coverage profile correction of shallow-depth circulating cell-free DNA sequencing via multidistance learning. In: Pacific symposium on biocomputing 2020. Stanford: World Scientific; 2019. p. 599–610.
66. Deshpande A, Walradt T, Hu Y, Koren A, Imielinski M. Robust foreground detection in somatic copy number data. bioRxiv. 2019;847681.

67. Gao GF, Oh C, Saksena G, Deng D, Westlake LC, Hill BA, et al. Tangent normalization for somatic copy-number inference in cancer genome analysis. Bioinformatics. 2022;38(20):4677–86.
68. Bonneel N, Van De Panne M, Paris S, Heidrich W. Displacement interpolation using Lagrangian mass transport. In: Proceedings of the 2011 SIGGRAPH Asia conference. New York. 2011. p. 1–12.
69. Courty N, Flamary R, Habrard A, Rakotomamonjy A. Joint distribution optimal transportation for domain adaptation. Adv Neural Inf Process Syst. 2017;30.
70. Raimondi D, Passemiers A, Verplaetse N, Corso M, Ferrero-Serrano Á, Nazzicari N, et al. Biologically meaningful genome interpretation models to address data underdetermination for the leaf and seed ionome prediction in Arabidopsis thaliana. Sci Rep. 2024;14(1):13188.
71. Benevenuta S, Fariselli P. On the upper bounds of the real-valued predictions. Bioinforma Biol Insights. 2019;13:1177932219871263.
72. Pedersen BS, Eyring K, De S, Yang IV, Schwartz DA. Fast and accurate alignment of long bisulfite-seq reads. 2014. arXiv preprint arXiv:1401.1129.
73. Liu T, Wang Z. SOV_refine: a further refined definition of segment overlap score and its significance for protein structure similarity. Source Code Biol Med. 2018;13(1):1–10.
74. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–30.
75. Lenaerts L, Brison N, Maggen C, Vancoillie L, Che H, Vandenberghe P, et al. Comprehensive genome-wide analysis of routine non-invasive test data allows cancer prediction: a single-center retrospective analysis of over 85,000 pregnancies. EClinicalMedicine. 2021;35:100856.
76. Buedts L, Wlodarska I, Finalet-Ferreiro J, Gheysens O, Dehaspe L, Tousseyn T, et al. The landscape of copy number variations in classical Hodgkin lymphoma: a joint KU Leuven and LYSA study on cell-free DNA. Blood Adv. 2021;5(7):1991–2002.
77. Bayindir B, Dehaspe L, Brison N, Brady P, Ardui S, Kammoun M, et al. Noninvasive prenatal testing using a novel analysis pipeline to screen for all autosomal fetal aneuploidies improves pregnancy management. Eur J Hum Genet. 2015;23(10):1286–93.
78. Vanderstichele A, Busschaert P, Smeets D, Landolfo C, Van Nieuwenhuysen E, Leunen K, et al. Chromosomal instability in cell-free DNA as a highly specific biomarker for detection of ovarian cancer in women with adnexal masses. Clin Cancer Res. 2017;23(9):2223–31.
79. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754–60.
80. Picard toolkit. Broad Institute. 2019. https://broadinstitute.github.io/picard/. Accessed 16 July 2019.
81. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20(9):1297–303.
82. Cleveland WS. Robust locally weighted regression and smoothing scatterplots. J Am Stat Assoc. 1979;74(368):829–36.
83. Cleveland WS, Grosse E, Shyu WM. Local regression models. In: Statistical models in S. Oxfordshire: Routledge; 2017. p. 309–76.
84. Seabold S, Perktold J. Statsmodels: econometric and statistical modeling with python. In: Proceedings of the 9th Python in science conference. vol. 57. Austin; 2010. p. 10–25080.
85. Huang J, Gretton A, Borgwardt K, Schölkopf B, Smola A. Correcting sample selection bias by unlabeled data. Adv Neural Inf Process Syst. 2006;19:601–8.
86. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: an imperative style, high-performance deep learning library. In: Advances in neural information processing systems 32. Red Hook: Curran Associates, Inc.; 2019. p. 8024–35.
87. Kingma DP, Ba J. Adam: a method for stochastic optimization. 2014. arXiv preprint arXiv:1412.6980.
88. Genevay A, Cuturi M, Peyré G, Bach F. Stochastic optimization for large-scale optimal transport. Adv Neural Inf Process Syst. 2016;29:3440–8.
89. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nature Methods. 2020;17:261–72. https://doi.org/10.1038/s41592-019-0686-2.
90. Fisher RA, et al. On the "probable error" of a coefficient of correlation deduced from a small sample. Metron. 1921;1:3–32.
91. Che H, Vermeesch J, Jatsenko T. Accurate multi-cancer detection using cell-free DNA shallow whole-genome sequencing data. 2022. https://www.ebi.ac.uk/biostudies/arrayexpress/studies/E-MTAB-10934.
92. Vanderstichele A, Busschaert P, Landolfo C, Olbrecht S, Coosemans A, Froyman W, et al. Low coverage WGS of plasma cfDNA. 2022. https://ega-archive.org/studies/EGAS00001005361.
93. Stanley KE, Jatsenko T, Tuveri S, Sudhakaran D, Lannoo L, Van Calsteren K, et al. WGS data of plasma cell free DNA. 2024. https://ega-archive.org/datasets/EGAD00001000856.
94. Passemiers A, Jatsenko T, Vermeesch JR. Cell-free DNA coverage and fragmentomic profiles produced under various sequencing protocols. 2024. https://doi.org/10.6084/m9.figshare.24459304.v1.
95. Passemiers A. DAGIP: a bias correction algorithm for cell-free DNA data. Zenodo. 2024. https://doi.org/10.5281/zenodo.14503340.
96. Passemiers A. DAGIP: a bias correction algorithm for cell-free DNA data. GitHub. 2024. https://github.com/JorisVermeeschLab/DAGIP.

Publisher's Note